Regresión

0. Análisis Exploratorio Inicial

Primero imprimí las columnas originales para ver exactamente cómo venían (en inglés) y dejar evidencia del estado inicial. Luego renombré todas las variables a español con un diccionario, pasando por ejemplo de Age a edad, Education_Level a nivel_educativo, Occupation a ocupacion e Income a ingreso, para estandarizar el dataset con nombres consistentes.

Después de eso también corregí la tipología de las variables: forcé las numéricas clave (edad, n_dependientes, exp_laboral, tam_hogar e ingreso) a formato numérico y convertí las categóricas (nivel_educativo, ocupacion, zona_residencia, estado_civil, situacion_laboral, tenencia_vivienda, tipo_vivienda, genero y transporte_principal) a tipo category. Finalmente volví a imprimir las columnas y los tipos para confirmar que tanto los nombres como los dtypes quedaron correctamente definidos.

1. Análisis Estadístico Inicial (Descriptivo)

1.1 Análisis Descriptivo Numérico

Con esto concluyo que las variables numéricas quedaron bien formateadas y completas: tengo 10,000 observaciones en todas y 0% de missing en edad, n_dependientes, exp_laboral, tam_hogar e ingreso. Los rangos son coherentes (edad 18–70, dependientes 0–5, experiencia 0–50, tamaño de hogar 1–7).

La señal más importante es **ingreso**: está fuertemente sesgado a la derecha, con media muy por encima de la mediana (816,838 vs 72,943) y una desviación estándar enorme, lo que sugiere outliers fuertes y alta dispersión. Además, los valores más frecuentes de ingreso se repiten muy poco (máximo 0.10%), así que no hay un “valor dominante”, mientras que en dependientes y tamaño de hogar sí hay concentraciones claras en pocos valores (por ejemplo 0–5 dependientes y hogares 1–7)

1.2 Análisis Descriptivo de Variables Categóricas

De aquí deduzco que las variables categóricas están “limpias” y controladas: todas tienen baja cardinalidad.

También se nota que varias variables tienen una categoría claramente dominante, lo cual me da pistas sobre el perfil de la muestra, zona_residencia es mayormente Urban (70%), estado_civil se concentra en Married (51%) y situacion_laboral en Full-time (50%). Eso sugiere que el dataset representa principalmente un segmento “formal/urbano”, por lo que

cualquier conclusión sobre ingresos podría estar más influenciada por ese grupo y menos por Rural o Self-employed, que son minoritarios pero no despreciables.

Por último, hay señales de “estructura” en categorías que probablemente capturen diferencias reales de ingreso: nivel_educativo está escalonado (High School/Bachelor/Master/Doctorate) con una cola pequeña en Doctorate (5%), ocupación está relativamente repartida pero con Healthcare y Technology liderando, y tenencia_vivienda (Own 60% vs Rent 40%) junto con tipo_vivienda (casi empate entre casa y departamento) sugieren variables con potencial fuerte para segmentar poder adquisitivo. En práctica, esperaría que educación y ocupación expliquen niveles, mientras zona_residencia y vivienda expliquen diferencias de costo de vida y patrimonio, y transporte_principal funcione como proxy de acceso y recursos.

2. Análisis de Calidad de los Datos (Diagnóstico)

2.1 Calidad de Variables Numéricas

Salió así porque al hacer el “coerce” antes y volverlo a hacer. no hubo nada que corregir, ya estaban en int64 y por eso dtype_original y dtype_post_coerce quedan iguales y además no aparecen missing. Los percentiles muestran que ingreso está fuertemente sesgado a la derecha, por eso el IQR marca muchos valores altos como outliers y te da 19.31% en ingreso.

2.2 Calidad de Variables Categóricas

Las variables categóricas están completamente limpias y consistentes porque no hay missing reales ni missing “codificados”, así que no necesito recodificar ni imputar nada. También deduzco que la estructura es estable y útil para modelar porque la cardinalidad es baja en todas, no existen categorías raras bajo 1%, y la dominancia es moderada a alta en algunas variables, especialmente zona_residencia con 70.37% Urban y tenencia_vivienda con 60.18% Own, lo que indica un perfil de muestra más urbano y propietario; en cambio ocupacion, nivel_educativo y transporte_principal están más repartidas, por lo que deberían aportar más variación explicativa sin riesgo de categorías marginales.

3. Preprocesamiento de Datos

3.1 Variables Numéricas

Verifiqué el missing antes y después de imputar y vi que no cambió nada, porque todas las variables numéricas ya tenían 0% de valores faltantes. Por eso, aunque apliqué SimpleImputer con mediana y dejé la evidencia de “antes vs después”, en la práctica la imputación no modificó los datos, solo confirmó que esa parte estaba limpia.

Luego hice la transformación logarítmica de ingreso y ahí sí se ve el efecto que buscaba, porque el sesgo bajó de 2.92 a 1.345. Eso me dice que ingreso tenía una cola derecha fuerte y que con el log reduzco esa asimetría, lo cual es útil para regresión.

También observo que al pasar por el imputer todas las columnas quedaron en float64, y esto pasa porque esa transformación devuelve números en formato float aunque originalmente fueran int. No es un problema, pero si quiero mantener enteros tendría que volver a convertir edad, n_dependientes, exp_laboral y tam_hogar al tipo original al final.

3.2 Variables Categóricas

Apliqué una limpieza básica de texto en todas las categóricas usando strip para eliminar espacios y asegurar que una misma etiqueta no aparezca duplicada por errores de formato, y confirmé que la cardinalidad quedó baja y estable en todas las variables, con rangos coherentes como 2 categorías en genero y tenencia_vivienda, 3 en zona_residencia y estado_civil, y 4 o 5 en nivel_educativo, transporte_principal y ocupacion. A partir de eso deduzco que el dataset no viene de texto libre desordenado, sino de un conjunto ya predefinido de categorías.

Luego verifiqué dos tipos de faltantes y ambos me salieron en cero, tanto los NaN reales como los “falsos missing” tipo ?, NA o vacío, así que al reemplazarlos y luego imputar con Missing en realidad no cambié datos. Como tampoco existen categorías raras por debajo de 1 por ciento, deduzco que no hay problemas de sparsidad ni niveles marginales que puedan romper una codificación posterior.

Finalmente, la auditoría de dominancia me muestra qué variables podrían sesgar el aprendizaje por tener una categoría muy dominante, especialmente zona_residencia con 70.37 por ciento Urban y tenencia_vivienda con 60.18 por ciento Own, mientras que ocupacion está más repartida y debería aportar variación útil. Con esto concluyo que las categóricas están listas para modelado y que si hago encoding después no voy a inflar demasiado el número de features ni voy a perder estabilidad por categorías raras.

4. Análisis Exploratorio de Datos (EDA Univariado) y Visualización

4.1 Variables Numéricas: Estadística Descriptiva y Visualización

En el histograma de edad se ve una distribución bastante pareja entre 18 y 70, lo que encaja con que los valores más frecuentes apenas llegan a 2.23 por ciento, así que no hay un pico dominante ni una concentración rara y un aporte de información al modelo.

Con respecto, n_dependientes y tam_hogar son variables discretas con masa repartida en pocos enteros y con frecuencias altas por valor, por ejemplo en n_dependientes los cinco valores más comunes están entre 16.42 y 17.45 por ciento cada uno. Esto sugiere que estas variables van a segmentar por saltos y no como una variable continua fina, por lo que es más realista esperar cambios por tramos que una relación suave.

En exp_laboral vs tam_hogar aparecen franjas horizontales perfectas porque tam_hogar solo toma valores enteros de 1 a 7, entonces el scatter no forma una nube continua sino bandas.

En n_dependientes vs ingreso veo columnas verticales porque n_dependientes también es discreta de 0 a 5, y dentro de cada columna el ingreso se dispersa casi igual de 0 hasta valores muy altos. Eso me sugiere que el número de dependientes por sí solo no ordena el ingreso de manera clara en escala original.

Además los outliers extremos de ingreso están presentes en todos los niveles de dependientes, por lo que esa cola derecha está tapando cualquier patrón fino y convendría mirar esta relación con ingreso_log o con medianas por bucket.

En edad vs n_dependientes vuelven a aparecer bandas horizontales y no se aprecia que la edad “empuje” a tener más o menos dependientes, porque para casi cualquier edad aparecen prácticamente todos los niveles de dependientes. Concluyó que estas tres relaciones, tal como están dibujadas en crudo, reflejan sobre todo la naturaleza discreta de las variables y la alta dispersión del ingreso, y que si quiero detectar estructura real necesito usar transformaciones o resúmenes robustos en lugar del scatter directo con ingreso.

4.2 Variables Categóricas: Frecuencias y Visualización

Dejé evidencia de cardinalidad por variable y el top de frecuencias con porcentajes, y además los countplots muestran visualmente la dominancia. La cardinalidad es baja en todas, entre 2 y 5 categorías, así que deduzco que la codificación futura no va a explotar en demasiadas columnas.

El conjunto está concentrado en ciertos perfiles por la dominancia del top 1 en varias variables. zona_residencia está fuertemente cargada a Urban con 70.37%, tenencia_vivienda favorece Own con 60.18%, y tanto estado_civil como situacion_laboral están inclinadas a Married 51.36% y Full time 50.04%. Eso sugiere que la muestra representa principalmente un segmento urbano y relativamente estable, por lo que el modelo va a aprender patrones más claros para ese grupo que para los segmentos menos frecuentes.

Veo dominancias claras, como zona_residencia muy cargada a Urban, tenencia_vivienda inclinada a Own, y situacion_laboral con Full time como la categoría más grande, lo que me indica que la muestra está compuesta mayormente por un perfil urbano y más estable.

Estas dominancias van a influir directamente en cómo se comporta ingreso cuando lo analice por grupos, porque la mayor parte del “promedio global” de ingreso va a estar determinado por Urban, Own y Full time, mientras que categorías menos frecuentes como Rural o Self employed aportarán menos peso aunque puedan tener patrones distintos. En ocupacion y nivel_educativo la distribución está más repartida y con jerarquía natural en educación, así que espero que ahí aparezcan diferencias más marcadas de ingreso entre categorías sin que una sola categoría opaque a las demás.

Observó variables más balanceadas que probablemente aporten señal sin sesgo extremo, como ocupacion donde el top 1 es 30.35% y el resto está bastante repartido, y transporte_principal donde Public transit 40.47% convive con Car 29.86%, Biking 19.40% y Walking 10.27%. En nivel_educativo la distribución tiene jerarquía natural con Bachelor 40.58% y Doctorate 5.01%, lo que me hace pensar que esa variable puede capturar diferencias estructurales en ingreso aunque el nivel más alto sea minoritario.

5. Análisis Exploratorio Bivariado (EDA): Variables Explicativas vs Target

5.1 Variables Numéricas vs Target

Tomamos ingreso como variable objetivo y armé el análisis solo con las variables numéricas explicativas, excluyendo ingreso e ingreso_log para no contaminar la evaluación. Para cada variable construí buckets por cuantiles y revisé ingreso dentro de cada rango usando boxplots, porque así detecto cambios de nivel, dispersión y la influencia de outliers sin asumir linealidad.

En edad, los promedios por bucket se mueven entre 763,698 y 858,499, pero la mediana se queda prácticamente plana alrededor de 72,900 a 73,216 en todos los rangos. De eso deduzco que la edad no está ordenando el ingreso típico, y que las diferencias que aparecen en la media se explican más por valores extremos dentro de ciertos grupos que por un cambio estructural del ingreso central. Esto cuadra con la etiqueta de tendencia no monotónica y con Spearman no significativo en buckets para la mediana.

En n_dependientes, la media sube en los buckets intermedios y luego vuelve a caer, y por eso también sale no monotónica en media. Sin embargo, la mediana sí muestra una subida gradual desde 72,664 hasta alrededor de 73,228 y se mantiene alta, y ahí Spearman por buckets para la mediana es 0.900 con p 0.0374. deduzco que hay una relación positiva leve en el ingreso típico cuando aumentan dependientes, pero es pequeña en magnitud y la media sigue siendo inestable por la cola derecha del ingreso.N_dependientes sí ordena el ingreso típico porque la línea de la mediana sube de forma bastante limpia conforme aumentan los dependientes, mientras que la media hace una panza y luego baja, lo que me indica que los outliers están inflando algunos buckets y por eso la media no es tan estable como la mediana.

En exp_laboral, la lectura es más clara porque tanto la media como la mediana caen conforme suben los buckets, y Spearman por buckets es -0.900 con p 0.0374 para ambas, dentro de esta muestra, más experiencia laboral está asociada a menor ingreso, y lo deduzco como un patrón real del centro de la distribución porque la mediana también baja de 73,562 a 72,156. Esto sugiere que el dataset no representa una trayectoria salarial típica, sino quizá una construcción donde exp_laboral está capturando otra cosa como antigüedad sin progresión o perfiles distintos.

En tam_hogar, la mediana es prácticamente plana y la media se mueve sin orden, con Spearman cero en mean y median, así que no hay señal monotónica. Deduzco que el tamaño del hogar no explica el ingreso típico y que las variaciones de la media por bucket son ruido inducido por outliers en ingreso, especialmente porque el último bucket vuelve a tener una media alta sin que la mediana cambie.

La única dirección clara y relativamente consistente es que a mayor exp_laboral, menor ingreso,edad y tam_hogar no muestran una dirección estable con ingreso, porque las medianas se mantienen casi planas y las medias se mueven sin orden por la cola derecha, y tanto Pearson como Spearman salen prácticamente en cero y el ANOVA no detecta diferencias entre buckets. Por eso, si tengo que priorizar, me quedo con exp_laboral como el único predictor numérico con señal clara, luego n_dependientes como señal secundaria muy suave, y dejo edad y tam_hogar como variables que probablemente aporten poco por sí solas en ingreso original y que requieren ingreso_log o enfoques robustos para evaluarlas mejor.

5.2 Variables Categóricas vs Target

El promedio de ingreso no sigue el orden típico de más educación más ingreso. High School tiene el promedio más alto 864618 y Doctorate el más bajo 622169, pero cuando miro la mediana todo queda mucho más parejo 73452 en High School, 72888 en Bachelor's, 72747 en Master's y 71346 en Doctorate. Eso me hace deducir que el promedio está siendo empujado por pocos ingresos extremadamente altos y que la comparación más estable es la mediana. Además Doctorate tiene solo 5.01% de la muestra, así que cualquier lectura ahí es más frágil por tamaño.

En ocupacion, el patrón por promedio sí marca diferencias claras entre grupos, con Education como el mayor promedio 914794 y Finance como el menor 701172. Sin embargo, las medianas están todas muy cerca del rango 72228 a 73428, lo que refuerza la misma idea de antes, la "persona típica" no cambia tanto entre categorías, pero hay colas largas que inflan el promedio en algunas ocupaciones. La dispersión también me confirma eso porque las desviaciones estándar son enormes en todas las categorías y en Finance el rango intercuartílico es el más bajo, lo que sugiere menos variabilidad en el centro de la distribución.

En zona_residencia, encuentro una señal interesante y contraintuitiva en el promedio, Rural es el más alto con 1042039, luego Suburban con 914476 y Urban el más bajo con 749564, pero las medianas siguen cercanas 74815, 73198 y 72664 respectivamente. Esto me lleva a deducir que en Rural hay pocos casos con ingresos muy altos que empujan el promedio, mientras que para la mayoría de personas el ingreso típico no difiere tanto por zona. Para el modelado.

Las medianas están muy pegadas entre sí mientras los promedios cambian bastante. Eso me dice que la "persona típica" dentro de cada grupo gana parecido, pero existen pocos casos con ingresos extremadamente altos que empujan el promedio y agrandan la dispersión. Por eso los boxplots se ven con colas largas y muchos puntos arriba, y por eso es más confiable mirar mediana e IQR para entender el centro real del grupo, y el promedio para detectar dónde se concentran esos casos extremos.

En tenencia_vivienda la diferencia es fuerte y no es solo ruido: Rent tiene mean 951062.95 y median 75328 frente a Own con mean 718882.27 y median 71921, y además el IQR de Rent es enorme 658541.75 mientras Own es muy chico 43371.00. Deduzco que los que alquilan están mezclando dos mundos a la vez, un bloque grande con ingresos "normales" y un bloque de ingresos muy altos que dispara promedio y dispersión, mientras que en Own la mayoría está mucho más concentrada y con menos variación. En tipo_vivienda pasa algo parecido pero más moderado: Apartment tiene mayor promedio 867675.93 y un IQR grande 378286, Single-family home baja a 788437.58 con IQR 248435 y Townhouse es el menor 743148.36 con IQR 145217, así que acá sí hay señal de segmentación por vivienda, pero de nuevo la mayor parte del salto viene por la cola alta más que por cambios gigantes en la mediana.

En situacion_laboral el grupo Part-time sorprende porque aparece con mean 899087.67 y median 73512, por encima de Full-time y Self-employed, y también con IQR alto 382830.75,

así que es como un grupo heterogéneo donde conviven muchos ingresos normales con algunos casos muy altos. En estado_civil la diferencia es pequeña en mediana y moderada en promedio, Married lidera el mean 833617.59 pero las medianas rondan 72900 en todos, entonces la señal es débil para predecir ingreso por sí sola. En genero prácticamente no hay señal, los promedios son muy parecidos y las medianas también. En transporte_principal las medias ordenan Biking 842073.46, Walking 825607.88, Public transit 817161.33 y Car 778564.13, pero las medianas están casi iguales alrededor de 72900, así que concluyo que transporte sirve más como proxy suave y probablemente se va a explicar mejor cuando lo cruce con zona, ocupación y vivienda, porque solo no está moviendo el centro del ingreso, solo la cola y la dispersión.