

UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMÁTICA



MINERÍA DE DATOS · PRE PROCESAMIENTO

Limpieza de datos con el objetivo empleado – empleador

Bitácora de limpieza de la Encuesta de Hogares 2024

Estudiante: Saul Mijael Choquehuanca Huanca
Docente: Lic. Menfy Morales Rios
Sigla: INF 311

La Paz – Bolivia
2026

Contenido

1	Introducción	2
2	Objetivo	2
2.1	Idea base y pregunta de minería	2
2.2	Objetivo general	2
2.3	Objetivos específicos	2
2.4	Variable objetivo y población	2
2.5	Atributos seleccionados	3
3	Metodología	4
4	Desarrollo	4
4.1	Paso 1. Auditoría y perfilado inicial	4
4.2	Paso 2. Normalización de estructura y formatos	6
4.3	Paso 3. Gestión de duplicados	7
4.4	Paso 4. Tratamiento de valores faltantes	7
4.5	Paso 5. Detección y tratamiento de valores atípicos	7
4.6	Paso 6. Validación de consistencia y reglas de negocio	9
5	Bitácora de limpieza de datos	11
6	Resultados	14
6.1	Calidad del dataset antes y después	14
6.2	Perfil del empleado y del empleador	15
6.3	Relevancia de los atributos	16
7	Conclusiones	17
8	Entregables	17
	Referencias	18

1 Introducción

La Encuesta de Hogares (EH) que el Instituto Nacional de Estadística (INE) levanta cada año es la principal fuente de información sobre las condiciones de vida y de trabajo de la población boliviana. La base de personas de la EH 2024 (archivo `persona.csv`) reúne 39 497 registros y 275 variables. Tal como se publica no puede entrar a un algoritmo de minería de datos: mezcla poblaciones distintas, guarda las categorías como números sin etiqueta, tiene columnas casi vacías y contiene algunos valores imposibles.

Este informe documenta el pre procesamiento de esa base con un objetivo concreto, construido sobre la idea **empleado–empleador**. El trabajo sigue la metodología de seis pasos de la asignatura y registra cada modificación en una **Bitácora de Limpieza de Datos** (Sección 5), de modo que cualquier persona pueda reproducir o auditar el resultado. Todo el proceso está programado en R.

Tabla 1: Ficha del trabajo.

Elemento	Detalle
Datos de entrada	<code>persona.csv</code> : 39 497 filas × 275 columnas (34 MB). Encuesta de Hogares 2024, Instituto Nacional de Estadística (Bolivia).
Herramientas	R version 4.5.3 [5], con los paquetes <code>data.table</code> 1.18.6.1, <code>dplyr</code> 1.2.1, <code>VIM</code> 7.3.1, <code>ggplot2</code> 4.0.3, <code>jsonlite</code> 2.0.0.
Código	<code>01_limpieza.R</code> (los seis pasos y la bitácora) y <code>02_resultados.R</code> (figuras y datos del tablero).
Dataset limpio	<code>empleado_empleador_limpio.csv</code> : 7 660 filas × 29 columnas, sin valores faltantes.
Bitácora	<code>bitacora_limpieza.csv</code> : 26 registros con los doce campos del formato.
Tablero	https://saulchoque.github.io/mineria-datos-empleado-empleador/
Repositorio	https://github.com/SaulChoque/mineria-datos-empleado-empleador

2 Objetivo

2.1 Idea base y pregunta de minería

En toda unidad económica hay quien contrata y quien es contratado. La EH registra esa posición en la pregunta «En esta ocupación usted trabaja como...» (`s04b_12`). La pregunta que orienta el trabajo es: **¿qué características permiten distinguir a un empleador de un empleado?**

2.2 Objetivo general

Obtener, a partir de la base de personas de la Encuesta de Hogares 2024, un conjunto de datos **limpio, consistente y documentado** de la relación empleado–empleador en Bolivia, que permita caracterizar y diferenciar a quienes trabajan como **empleados** (obreros o empleados asalariados) de quienes lo hacen como **empleadores** (patrones o socios que contratan personal), como insumo para un modelo de clasificación del rol laboral.

2.3 Objetivos específicos

1. Delimitar la población objetivo y construir la variable objetivo `rol` (Empleado / Empleador).
2. Seleccionar los atributos que describen la relación empleado–empleador: demografía, educación, ocupación, unidad económica, jornada, ingreso y seguridad social.
3. Aplicar los seis pasos de la metodología de pre procesamiento, con las técnicas que corresponden a cada problema encontrado.
4. Registrar cada cambio en la bitácora de limpieza y publicar los resultados en un tablero.

2.4 Variable objetivo y población

De las ocho categorías ocupacionales solo tres forman una relación empleado–empleador dentro de una unidad económica (Tabla 2 y Figura 1). Se excluyen:

- **Cuenta propia, cooperativistas, familiares y aprendices sin remuneración:** no contratan ni son contratados por un salario.
- **Empleadas y empleados del hogar:** son asalariados, pero su empleador es un hogar y no una unidad económica; a ellos el cuestionario no les pregunta la administración ni el tamaño de la empresa, que son atributos centrales del estudio.
- **Menores de 14 años:** el INE define la población en edad de trabajar desde esa edad.

Tabla 2: Categoría ocupacional de las personas ocupadas en el dataset crudo y su destino.

Código	Categoría ocupacional (s04b_12)	Registros	Destino
1	Obrero/Empleado	6 951	rol = Empleado
2	Empleador/a socio con salario	55	rol = Empleador
3	Trabajador/a por cuenta propia	8 709	Fuera del objetivo
4	Empleador/a o socio/a sin salario	668	rol = Empleador
5	Cooperativista de producción	51	Fuera del objetivo
6	Familiar sin remuneración	2 633	Fuera del objetivo
7	Aprendiz sin remuneración	48	Fuera del objetivo
8	Empleada/o del hogar	364	Fuera del objetivo

Categoría ocupacional de las personas ocupadas (dataset crudo)

Solo tres de las ocho categorías forman una relación empleado-empleador

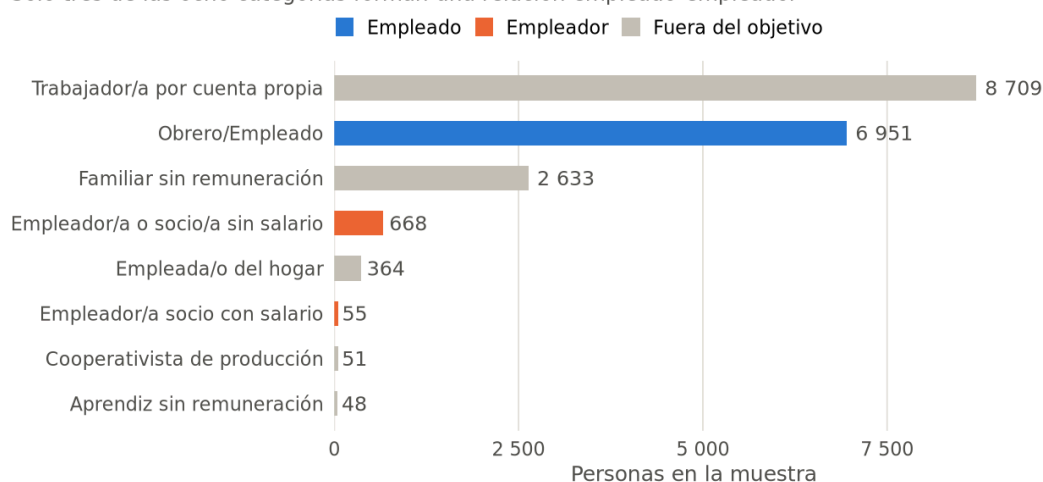


Figura 1: Las tres categorías que entran al estudio frente a las cinco excluidas.

2.5 Atributos seleccionados

De las 275 columnas se conservaron las 33 que describen a la persona, a su unidad económica y a su vínculo laboral (Tabla 3). Salud, fecundidad, migración, uso de TIC, bonos e ingresos no laborales no aportan a la pregunta planteada.

Tabla 3: Atributos del crudo seleccionados según el objetivo.

Grupo	Variables originales	Aporte al objetivo
Identificación y diseño	folio, nro, upm, estrato, factor	Clave de la persona y ponderador para estimar población.
Ubicación	depto, area	Contexto geográfico.
Demografía	s01a_02, s01a_03, s01a_04a, s01a_04b, s01a_04c, s01a_05, s01a_10	Sexo, edad, fecha de nacimiento, jefatura de hogar y estado civil.
Educación	niv_ed, aestudio	Capital humano.
Ocupación	s04b_12, cob_op, caeb_op	Variable objetivo, grupo ocupacional y rama de actividad.
Unidad económica	s04b_13, s04b_14, s04b_11aa, s04b_11ab	Sector, tamaño y antigüedad.

Grupo	Variables originales	Aporte al objetivo
Jornada	s04b_15, s04b_16aa, s04b_16ab, phrs, shrs, s04e_25	Días y horas trabajadas; segunda ocupación.
Ingreso	yprilab, yseclab	Ingreso laboral mensual.
Seguridad social	s04f_35, s04f_36	Afiliación y aporte a la AFP.

3 Metodología

Se aplicó la metodología para el pre procesamiento de datos de la asignatura [2], que ordena la limpieza en seis pasos con retorno (Figura 2): si un paso genera un problema nuevo, se vuelve a un paso anterior.

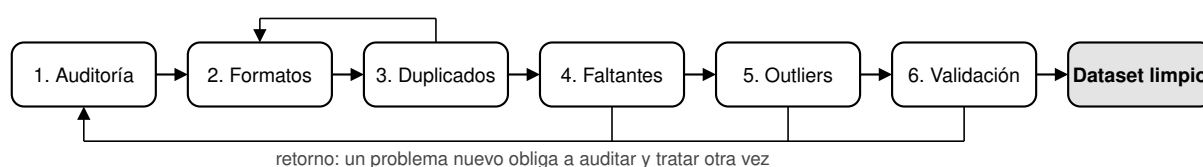


Figura 2: Metodología de pre procesamiento en seis pasos con retorno.

Tabla 4: Pasos de la metodología y técnicas utilizadas.

Paso	Técnicas aplicadas en este trabajo	Registros en bitácora
1. Auditoría y perfilado	Tipos de dato, estadísticas descriptivas, faltantes por columna, cardinalidad y cruces de verificación; delimitación de población y selección de atributos.	4
2. Estructura y formatos	Renombrado, trimming y mayúsculas, decodificación de categorías, fecha ISO-8601, unidad única, precisión decimal y construcción de atributos.	7
3. Duplicados	Duplicados exactos y parciales con regla de decisión documentada.	2
4. Valores faltantes	Eliminación de columnas con más de 50 %, imputación por regla lógica, imputación KNN e imputación por mediana.	4
5. Valores atípicos	Detección por IQR y Z-score; anulación de errores, winsorización, transformación logarítmica y discretización.	6
6. Validación	11 reglas de negocio ejecutadas en ciclo hasta no encontrar violaciones; auditoría final.	3

El retorno ocurrió en tres momentos: al anular una jornada imposible en el paso 5 se creó un valor faltante que obligó a volver al paso 4; la validación del paso 6 necesitó dos pasadas; y el proceso se cerró repitiendo la auditoría del paso 1 sobre el resultado.

Cada acción se anotó en la bitácora con los doce campos del formato: ID / paso, fecha y hora, responsable, dataset, atributo, tipo de problema, metodología, acción realizada, criterio o justificación, filas afectadas, impacto en el dataset y código. El script la llena automáticamente con la función `registrar()`, de modo que las cifras de la bitácora son las que realmente produjo la ejecución.

4 Desarrollo

4.1 Paso 1. Auditoría y perfilado inicial

Antes de modificar cualquier dato se perfilaron las 275 columnas: tipo, mínimo, máximo, media, mediana, número de faltantes y de valores distintos (registro 1 de la bitácora).

Tipos de dato. 233 columnas son enteras, 26 decimales y 16 de texto. Casi todas las enteras son en realidad categorías codificadas.

Faltantes. 179 de las 275 columnas están vacías en más de la mitad de los registros (Figura 3). No es descuido de la encuesta sino consecuencia de los saltos del cuestionario: las preguntas de empleo solo se hacen a quien trabaja, las de fecundidad solo a mujeres de cierta edad, etc.

Las 275 columnas del crudo según su porcentaje de valores faltantes

Dos de cada tres columnas están vacías en más de la mitad de los registros

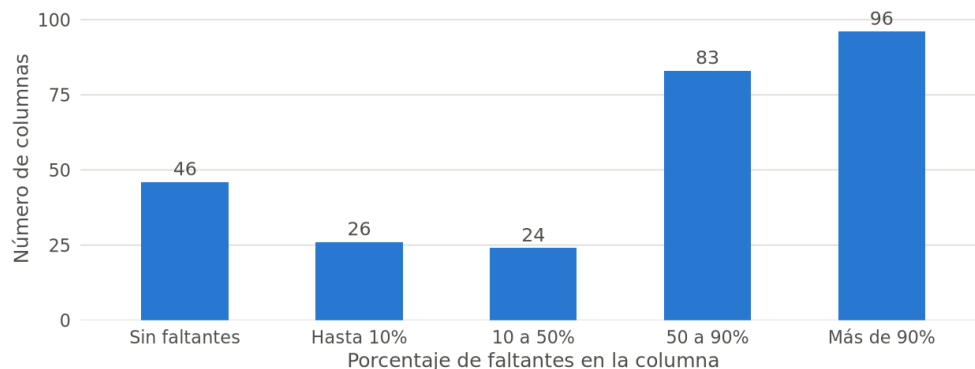


Figura 3: Columnas del dataset crudo según su porcentaje de valores faltantes.

Cardinalidad y significado de los códigos. El diccionario entregado describe las variables por grupos y en dos casos no coincide con los datos: s01a_02 figura como parentesco y s01a_03 como sexo, pero la primera solo toma los valores 1 y 2 y la segunda va de 0 a 98. Las etiquetas se tomaron entonces del diccionario oficial del INE [1] y se comprobaron con cruces: la edad media por código de estado civil es máxima en el código 5 (68 años, viudos) y mínima en el 6 (24 años, solteros), y todas las personas con categoría 8 (empleadas del hogar) caen en la rama 19 (hogares privados).

El resultado del paso es la lista de anomalías de la Tabla 5.

Tabla 5: Lista inicial de anomalías.

N.º	Anomalía	Evidencia en los datos	Paso que la trata
A01	Exceso de atributos para el objetivo	275 columnas; 179 (65%) tienen más de 50% de faltantes.	1 (delimitación)
A02	Población fuera del objetivo	De 39 497 personas solo 19 479 están ocupadas y solo 7 674 son obrero/empleador o empleador; 11 de ellas tienen menos de 14 años.	1 (delimitación)
A03	Clases desbalanceadas en la variable objetivo	6 951 obreros/empleados frente a 723 empleadores (razón 9.6 a 1).	Se documenta (modelado)
A04	Catóricas guardadas como enteros sin etiqueta	233 columnas enteras y 26 decimales; sexo, departamento, nivel educativo, categoría ocupacional, etc. son códigos numéricos.	2
A05	Diccionario entregado no coincide con los datos	s01a_02 toma valores {1, 2} (es sexo, no parentesco) y s01a_03 va de 0 a 98 (es edad, no sexo); estado civil viene codificado 1-6 y no 1,2,4,5,6,7 como el cuestionario.	2
A06	Fecha de nacimiento partida en 3 columnas	s01a_04a (día), s01a_04b (mes) y s01a_04c (año) en columnas separadas; no hay una fecha ISO-8601.	2
A07	Antigüedad laboral en unidades mezcladas	s04b_11aa se expresa en la unidad de s04b_11ab: 223 en semanas, 2 112 en meses y 17 144 en años (entre ocupados).	2
A08	Artefactos de precisión decimal en ingresos	7 887 valores de yprilab no son múltiplos de 0.01 (p. ej. 4524.8500976562): residuo de conversión float de 32 bits.	2
A09	Vacíos y códigos centinela	14 de 16 columnas de texto usan cadena vacía en vez de NA; valores centinela 999999 en s04c_19aa: 1.	1 (delimitación)
A10	Faltantes estructurales por saltos del cuestionario	s04b_13 (administración) no se pregunta a 668 empleadores sin salario; s04f_36 (aporta a AFP) solo a afiliados; horas e ingreso de ocupación secundaria vacíos en 93% de los ocupados.	2 y 4
A11	Ingreso laboral faltante en ocupados	3 registros de la población objetivo sin yprilab.	4

N.º	Anomalía	Evidencia en los datos	Paso que la trata
A12	Jornadas laborales imposibles o inverosímiles	phrs máximo = 171.5 h/semana, cuando la semana solo tiene 168 h; registros que superan 168 h: 1; ocupados con jornadas de 24 h diarias: 58.	5
A13	Ingreso laboral muy asimétrico con valores extremos	yprilab: mediana 3 300, máximo 54 000 Bs/mes, asimetría 3.3.	5
A14	Inconsistencias con la relación empleado- -empleador	50 obreros/empleados y 3 empleadores declaran que en su empresa trabaja 1 sola persona; 3 empleadores figuran en la administración pública.	6

Delimitación según el objetivo. Las tres primeras anomalías se resuelven acotando el dataset (Figura 4): se conservan las categorías 1, 2 y 4 de s04b_12 (registro 2, 31 823 filas eliminadas), las personas de 14 años o más (registro 3, 11 filas) y las 33 columnas de la Tabla 3 (registro 4). Estas filas y columnas no se descartan por tener errores, sino porque no pertenecen a la pregunta.

De la encuesta completa a la población objetivo

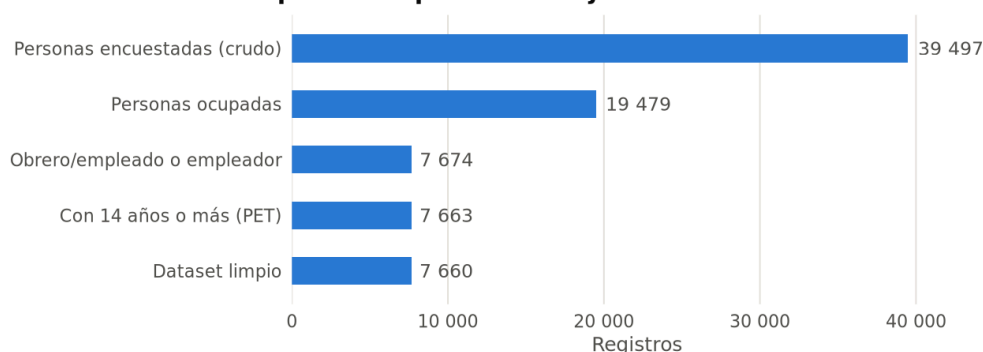


Figura 4: Registros que quedan tras cada filtro de delimitación.

4.2 Paso 2. Normalización de estructura y formatos

Sobre el subconjunto de trabajo se aplicaron siete acciones (registros 5 a 11), ilustradas con ejemplos reales en la Tabla 6:

- **Nombres:** los códigos del cuestionario pasan a nombres descriptivos en snake_case.
- **Texto:** trimws() y toupper() en los identificadores; se crea la clave única id_persona.
- **Categorías:** 11 columnas enteras se convierten en factores con etiqueta; ningún código quedó fuera de catálogo.
- **Estructura:** se construyen rol (variable objetivo), sector, jefe_hogar y afp. Esta última fusiona la afiliación y el aporte a la AFP: como el aporte solo se pregunta a los afiliados, el vacío se convierte en la categoría explícita «No afiliado».
- **Fechas:** día, mes y año se unen en fecha_nacimiento con formato ISO-8601; ninguna fecha resultó inexistente.
- **Unidades:** la antigüedad, declarada en semanas, meses o años, se lleva a años (1 499 registros convertidos).
- **Precisión:** los montos se redondean a centavos (3 692 registros con residuos de conversión).

Tabla 6: Ejemplos de normalización de formato.

Técnica	Antes (crudo)	Después
Renombrado	s04b_12, yprilab, phrs	categoría_ocupacional, ingreso_laboral, horas_semana
Decodificación	s01a_02 = 2	sexo = Mujer
Decodificación	depto = 7	departamento = Santa Cruz
Fecha ISO-8601	s01a_04a = 5, s01a_04b = 2, s01a_04c = 1975	fecha_nacimiento = 1975-02-05

Técnica	Antes (crudo)	Después
Unidad única	s04b_11aa = 6, s04b_11ab = 4 (meses)	antiguedad_anios = 0.5
Precisión decimal	yprilab = 4182.6665039062	ingreso_laboral = 4182.67
Construcción de atributo	s04f_35 = 2, s04f_36 = (vacío)	afp = No afiliado
Construcción de atributo	s04b_12 = 4	rol = Empleador
Clave única	folio = 111-000187-A-0071, nro = 1	id_persona = 111-000187-A-0071-01

4.3 Paso 3. Gestión de duplicados

No hay filas idénticas ni identificadores repetidos (registros 12 y 13). La búsqueda de duplicados parciales sí encontró cuatro registros, en dos hogares, iguales en todas las columnas seleccionadas salvo el número de persona. La regla de decisión fue no eliminar sin revisar: cada par se comparó con su registro completo de 275 columnas y resultó que difieren en 3 y 5 respuestas (gasto en salud, idiomas, uso de TIC). Comparten fecha de nacimiento y son hijos del mismo hogar: son hermanos gemelos que trabajan juntos, no una persona digitada dos veces. Se conservaron.

4.4 Paso 4. Tratamiento de valores faltantes

Tras el paso 2 solo cuatro columnas tenían vacíos (Tabla 7). Cada una se trató según su proporción y, sobre todo, según la causa del vacío.

Tabla 7: Valores faltantes en el subconjunto de trabajo y su tratamiento.

Columna	Faltantes	%	Causa	Tratamiento
ingreso_laboral_sec	7 347	95.88	Estructural (solo quien tiene segunda ocupación)	Columna eliminada (más de 50%)
horas_semana_sec	7 255	94.68	Estructural (solo quien tiene segunda ocupación)	Columna eliminada (más de 50%)
sector	668	8.72	Estructural (no se pregunta al empleador sin salario)	Regla lógica: sector = Privado
ingreso_laboral	3	0.04	Dato no declarado	Imputación KNN (k = 5)

- **Imputación por regla lógica (registro 14).** La administración de la empresa no se pregunta al empleador que no recibe salario. No es un dato perdido: quien dirige su propio negocio pertenece al sector privado. 668 registros.
- **Eliminación de columnas (registro 15).** Las horas y el ingreso de la ocupación secundaria superan el 50 % de vacíos, el umbral de la metodología. Lo rescatable, si la persona tiene o no una segunda ocupación, ya está completo en ocupacion_secundaria.
- **Imputación avanzada KNN (registro 16).** Los 3 ingresos no declarados se estimaron con VIM::knn [3]: para cada registro se buscan los 5 más parecidos según la distancia de Gower sobre 14 atributos (rol, sexo, edad, educación, ocupación, rama, sector, tamaño, antigüedad, horas, departamento, área y AFP) y se asigna la mediana de sus ingresos, que resiste mejor la asimetría que la media.

Al terminar el paso no quedó ningún valor faltante.

4.5 Paso 5. Detección y tratamiento de valores atípicos

Las siete variables numéricas se examinaron con los dos criterios de la metodología, rango intercuartílico y Z-score (Tabla 8), que son también los habituales en la literatura [4]. Los criterios marcan candidatos; la decisión se tomó variable por variable.

Tabla 8: Detección de valores atípicos antes del tratamiento.

Variable	Mín.	Mediana	Máx.	Asimetría	Fuera de IQR	Z > 3	Decisión
edad	14	37	84	0.52	21	18	Se conservan (valores reales)
años_estudio	0	14	23	-0.91	330	66	Se conservan (valores reales)
antigüedad_años	0.02	4	60	1.94	469	149	Se conservan (valores reales)
tam_empresa	1	7	3 000	12.87	793	58	Discretización en 4 tramos
días_semana	1	6	7	-1.44	469	204	Se conservan (valores reales)
horas_semana	2	45	171.5	0.57	1 003	66	Error > 168 h anulado e imputado; capado en Z = 3
ingreso_laboral	83.33	3 300	54 000	3.3	433	147	Capado P1-P99 por rol + logaritmo

- **Error confirmado (registros 17 y 18).** Un registro declaraba 171.5 horas de trabajo semanales, cuando la semana tiene 168. El valor se anuló, no la fila, y se volvió al paso 4 para imputarlo con la mediana de su rol.
- **Winsorización de horas (registro 19).** El IQR marcaba 1 003 registros, incluidas jornadas parciales y de 60 a 70 horas que son reales. Se usó el límite $Z = 3$: las 65 jornadas superiores a 92.3 horas se reemplazaron por ese valor (Figura 7).
- **Winsorización del ingreso (registro 20).** Los ingresos extremos son plausibles, pero distorsionan medias y distancias. Se caparon en los percentiles 1 y 99 de cada rol (Empleado: 520 a 14 038 Bs; Empleador: 323 a 22 258 Bs), 146 registros. Un límite común habría borrado la diferencia entre roles que se quiere estudiar (Figura 5).
- **Transformación logarítmica (registro 21).** Se añadió `log_ingreso_laboral`; la asimetría del ingreso baja de 3.3 a un valor cercano a cero (Figura 6).
- **Discretización (registro 22).** El tamaño de la unidad económica llega a 3 000 personas; son instituciones reales. En lugar de capar se añadió `tam_empresa_cat` (micro, pequeña, mediana, grande).
- **Sin tratamiento (registro 23).** Los valores marcados en edad, años de estudio, antigüedad y días trabajados son casos reales que el modelo debe conocer.

Ingreso laboral mensual antes y después del tratamiento de atípicos

Tramos de 1 000 Bs; el último tramo agrupa los ingresos mayores a 26 000 Bs

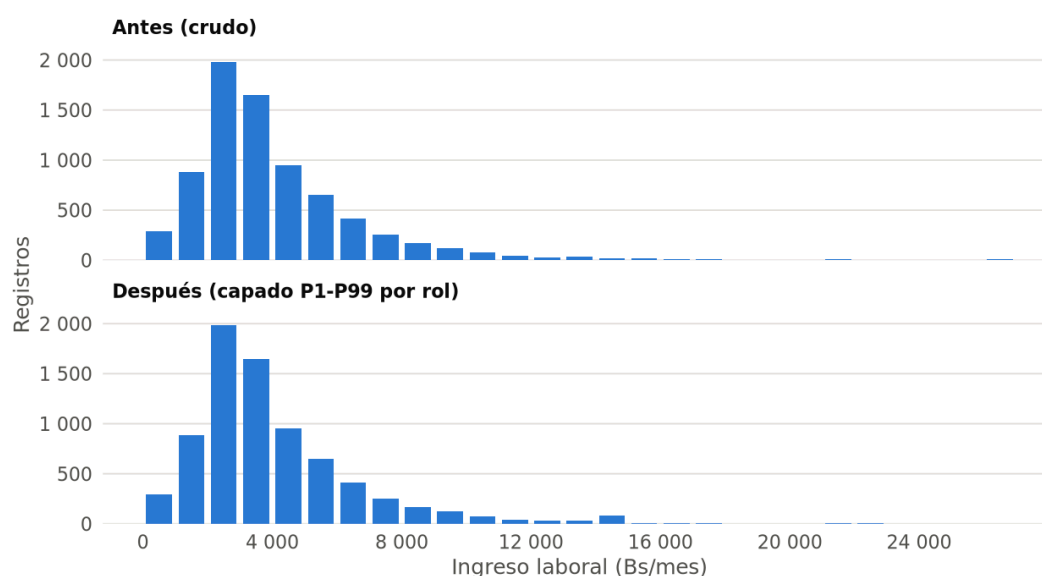


Figura 5: Ingreso laboral antes y después de la winsorización por rol.

Ingreso laboral tras la transformación logarítmica

La asimetría pasa de 3.30 (bolivianos) a -0.31 (logaritmo)

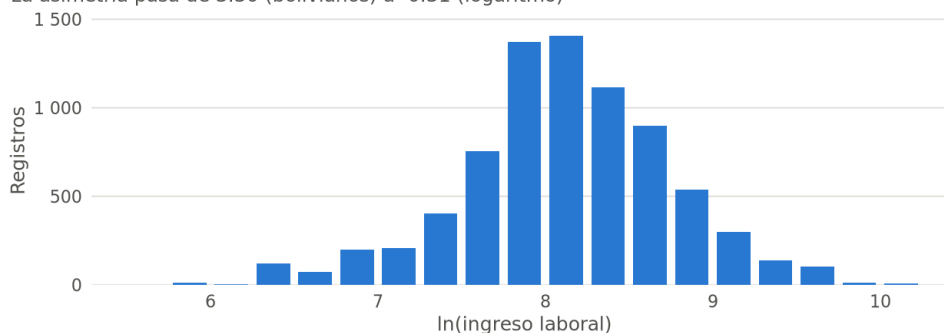


Figura 6: Distribución del logaritmo del ingreso laboral.

Horas trabajadas por semana antes y después

Una semana tiene 168 horas; el límite de capado quedó en 92.3 h

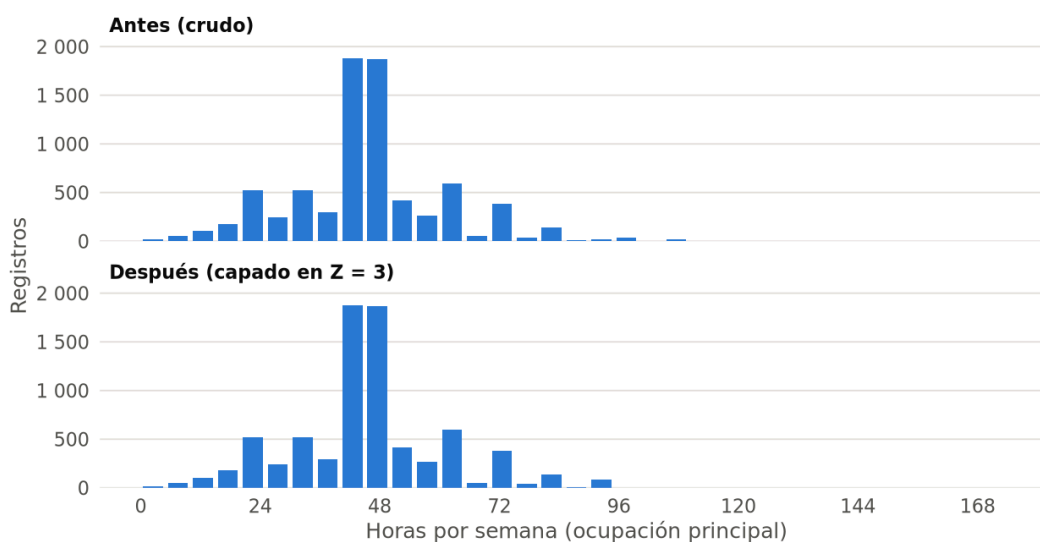


Figura 7: Horas semanales antes y después del tratamiento.

4.6 Paso 6. Validación de consistencia y reglas de negocio

Se definieron 11 reglas de negocio (Tabla 9) que se ejecutan en ciclo: si alguna falla se corrige y se vuelve a validar todo.

Tabla 9: Reglas de negocio y violaciones encontradas en cada iteración.

Regla	Debe cumplirse que...	Violaciones, iteración 1	Violaciones, iteración 2
R01	La edad está entre 14 y 120 años	0	0
R02	La fecha de nacimiento es anterior a la encuesta y coincide con la edad (± 1 año)	0	0
R03	La antigüedad no supera a la edad menos 7 años	0	0
R04	Los años de estudio no superan a la edad menos 4 años	0	0
R05	Se trabaja entre 1 y 7 días y entre 1 y 168 horas por semana	0	0
R06	El ingreso laboral es mayor que cero	0	0
R07	La unidad económica tiene al menos 2 personas (empleado + empleador)	53	0
R08	Un empleador no pertenece al sector público	3	0
R09	Nivel educativo 'Ninguno' equivale a 0 años de estudio	0	0
R10	El rol coincide con la categoría ocupacional y el id es único	0	0
R11	Las columnas derivadas coinciden con su origen (log e intervalo de tamaño)	0	0

La primera pasada encontró dos problemas que ninguna técnica estadística había señalado, porque nacen de la propia idea empleado–empleador:

- **R07 (registro 24).** 53 personas declararon que en su empresa trabaja una sola persona. Una relación empleado–empleador necesita al menos dos; el valor se corrigió al mínimo lógico (2) sin eliminar la fila.
- **R08 (registro 25).** 3 empleadores figuraban en la administración pública. El Estado no tiene patrones ni socios: la categoría ocupacional o el sector están mal registrados y no hay forma de saber cuál. Como `rol` es la variable objetivo, una etiqueta dudosa contamina el modelo y las filas se eliminaron.

La segunda pasada no encontró violaciones. El proceso se cerró repitiendo el perfilado del paso 1 sobre el resultado: cero faltantes, cero duplicados (registro 26).

5 Bitácora de limpieza de datos

Registro de auditoría generado por 01_limpieza.R durante la ejecución. Dataset de origen: persona.csv (EH 2024). El campo «Impacto en dataset» indica las dimensiones antes y después de cada cambio.

ID / Paso (#)	Fecha / Hora	Responsable	Dataset	Atributo / Variable	Tipo de Problema	Metodología	Acción Realizada	Criterio / Justificación	Filas Afectadas (N)	Impacto en Dataset	Código / Script
1	2026-10-07 11:30:19	Saul Mijael Choquehuanca Huanca	persona.csv (EH 2024)	Todas (275 columnas)	Estado del dataset desconocido	1. Auditoría y perfilado	Perfilado completo sin modificar datos: tipos, mínimos, máximos, medias, faltantes por columna y cardinalidad de categóricas.	Antes de modificar cualquier dato se evalúa el estado general. Resultado: lista de 14 anomalías (A01-A14) que guían los pasos siguientes.	0	Antes: 39 497 filas x 275 col. Después: 39 497 filas x 275 col.	perfilar(crudo); table(); tapply()
2	2026-10-07 11:30:19	Saul Mijael Choquehuanca Huanca	persona.csv (EH 2024)	s04b_12 (categoría ocupacional)	Registros fuera de la población objetivo	1. Auditoría y perfilado	Eliminación de filas: se conservan solo obrero/empleado (1) y empleador/socio (2 y 4).	El objetivo estudia la relación empleado- empleador. Inactivos, desocupados, cuenta propia, cooperativistas, familiares/aprendices sin pago y empleadas del hogar no forman parte de esa relación dentro de una unidad económica.	31 823	Antes: 39 497 filas x 275 col. Después: 7 674 filas x 275 col.	filter(s04b_12 %in% c(1, 2, 4))
3	2026-10-07 11:30:19	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	s01a_03 (edad)	Registros fuera de la población objetivo	1. Auditoría y perfilado	Eliminación de filas con edad menor a 14 años.	Regla de negocio: el INE define la Población en Edad de Trabajar (PET) desde los 14 años; el trabajo infantil no es una relación laboral formal comparable.	11	Antes: 7 674 filas x 275 col. Después: 7 663 filas x 275 col.	filter(s01a_03 >= 14)
4	2026-10-07 11:30:19	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	242 columnas descartadas	Atributos irrelevantes para el objetivo	1. Auditoría y perfilado	Selección de atributos: se conservan 33 columnas (identificación, demografía, educación, ocupación, jornada, ingreso laboral y AFP).	Reducción de dimensionalidad orientada al objetivo: salud, fecundidad, TIC, migración, bonos y texto libre no describen la relación empleado-empleador; además concentran los vacíos y códigos centinela detectados.	7 663	Antes: 7 663 filas x 275 col. Después: 7 663 filas x 33 col.	select(all_of(vars_ objetivo))
5	2026-10-07 11:30:19	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	33 columnas	Inconsistencia de formato (nombres crípticos)	2. Normalización de estructura y formatos	Renombrado de columnas a nombres descriptivos en snake_case (p. ej. s04b_12 -> categoria_ ocupacional, yprilab -> ingreso_laboral).	Los códigos del cuestionario (s04b_14, phrs...) no son legibles y el diccionario entregado no coincide con algunos de ellos (A05); los nombres se fijaron con el diccionario oficial del INE.	7 663	Antes: 7 663 filas x 33 col. Después: 7 663 filas x 33 col.	names(trabajo) <- nombres[names(trabajo)]
6	2026-10-07 11:30:19	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	folio, upm, estrato, id_ persona	Inconsistencia de formato (texto e identificador)	2. Normalización de estructura y formatos	trimws() + toupper() en folio y upm (0 valores cambiaron); estrato pasa de entero a texto; se crea la clave id_ persona = folio + nro_persona.	Los identificadores deben compararse sin espacios ni diferencias de mayúsculas; estrato es un código, no una magnitud; la persona se identifica por hogar + número de orden y sin clave única no se pueden buscar duplicados.	7 663	Antes: 7 663 filas x 33 col. Después: 7 663 filas x 34 col.	mutate(folio = toupper(trimws(folio)), id_persona = sprintf('%s-%02d', folio, nro_persona))
7	2026-10-07 11:30:19	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	sexo, area, departamento, estado_civil, nivel_ educativo, categoria_ ocupacional, grupo_ ocupacional, actividad_economica, ocupacion_ secundaria, afiliado_ afp, aporta_afp	Inconsistencia de formato (códigos numéricos sin etiqueta)	2. Normalización de estructura y formatos	Recodificación de 11 columnas enteras a factor con etiquetas del diccionario oficial; 0 códigos quedaron fuera de catálogo.	Un entero se interpreta como magnitud (Pando = 9 no es 'más' que La Paz = 2). Las etiquetas se validaron por cardinalidad y con cruces (edad media por estado civil, rama de las empleadas del hogar).	7 663	Antes: 7 663 filas x 34 col. Después: 7 663 filas x 34 col.	mutate(sexo = factor(sexo, levels = 1:2, labels = c('Hombre', 'Mujer')), ...)
8	2026-10-07 11:30:19	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	rol (objetivo), sector, jefe_hogar, afp	Estructura no orientada al objetivo	2. Normalización de estructura y formatos	Construcción de atributos: rol (Empleado/ Empleador) desde categoria_ ocupacional; sector (Público/Privado/ONG) agrupando los 6 códigos de administración; jefe_ hogar desde parentesco; afp fusiona afiliado_afp y aporta_afp en 3 categorías.	rol es la variable objetivo. aporta_afp solo se pregunta a afiliados (salto del cuestionario): fusionarla con afiliado_afp convierte ese vacío estructural en la categoría explícita 'No afiliado'.	7 663	Antes: 7 663 filas x 34 col. Después: 7 663 filas x 34 col.	mutate(rol = ifelse(categoria_ ocupacional == 'Obrero/Empleado', 'Empleado', 'Empleador'), afp = case_when(...))
9	2026-10-07 11:30:19	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	fecha_nacimiento (s01a_04a, s01a_04b, s01a_04c)	Inconsistencia de formato (fecha fragmentada)	2. Normalización de estructura y formatos	Se unen día, mes y año en una sola columna Date con formato ISO-8601; 0 fechas resultaron inexistentes en el calendario.	Tres enteros sueltos no son una fecha: no se pueden ordenar ni restar. as.Date() además valida el calendario (un 30 de febrero quedaría como NA).	7 663	Antes: 7 663 filas x 34 col. Después: 7 663 filas x 32 col.	as.Date(sprintf('%04d-%02d- %02d', nac_anio, nac_mes, nac_dia))
10	2026-10-07 11:30:19	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	antigüedad_anios (s04b_11aa, s04b_11ab); horas_dia (s04b_16aa, s04b_16ab)	Inconsistencia de formato (unidades mezcladas)	2. Normalización de estructura y formatos	Conversión a una sola unidad: antigüedad en años (semanas/52, meses/12) y jornada diaria en horas decimales (horas + minutos/60).	El valor 6 puede ser 6 semanas, 6 meses o 6 años según otra columna; sin unificar, la variable no es comparable entre registros.	1 499	Antes: 7 663 filas x 32 col. Después: 7 663 filas x 30 col.	case_when(antig_unidad == 8 ~ antig_cantidad, antig_ unidad == 4 ~ antig_ cantidad / 12, antig_unidad == 2 ~ antig_cantidad / 52)

ID / Paso (#)	Fecha / Hora	Responsable	Dataset	Atributo / Variable	Tipo de Problema	Metodología	Acción Realizada	Criterio / Justificación	Filas Afectadas (N)	Impacto en Dataset	Código / Script
11	2026-10-07 11:30:19	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	ingreso_laboral, ingreso_laboral_sec	Inconsistencia de formato (precisión decimal)	2. Normalización de estructura y formatos	Redondeo de los montos a 2 decimales (centavos); edad, años_estudio y tam_empresa se fijan como enteros.	Valores como 4182.6665039062 son residuos de una conversión float de 32 bits; un monto en bolivianos solo tiene sentido hasta el centavo.	3 692	Antes: 7 663 filas x 30 col. Después: 7 663 filas x 30 col.	mutate(across(all_of(col_ingreso), ~ round(.x, 2)))
12	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	Todas	Duplicado exacto	3. Gestión de duplicados	Búsqueda de filas idénticas en todas las columnas; regla de decisión: conservar la primera aparición.	Una fila repetida cuenta dos veces a la misma persona y sesga conteos y modelos. No se encontró ninguna.	0	Antes: 7 663 filas x 30 col. Después: 7 663 filas x 30 col.	trabajo[!duplicated(trabajo),]
13	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	id_persona; resto de columnas	Duplicado parcial	3. Gestión de duplicados	(a) Mismo id_persona repetido: 0 casos. (b) Mismo hogar e iguales en las 28 columnas restantes: 4 registros (2 pares). Cada par se comparó con su registro completo de 275 columnas: difieren en 3 a 5 respuestas, por lo que se conservan.	Regla de decisión: solo se elimina si el registro completo también es idéntico. Los pares comparten fecha de nacimiento y son hijos del mismo hogar con respuestas distintas en salud, idiomas y uso de TIC: son hermanos gemelos que trabajan juntos, no una persona digitada dos veces.	0	Antes: 7 663 filas x 30 col. Después: 7 663 filas x 30 col.	duplicated(trabajo[, sin_id]); comparar con crudo[folio + nro,] antes de eliminar
14	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	sector	Valor nulo (faltante estructural)	4. Valores faltantes	Imputación categórica por regla lógica: sector = 'Privado' en los empleadores a quienes no se hizo la pregunta.	La pregunta de administración no se aplica a 'empleado o socio que no recibe salario'. No es un dato perdido: quien dirige su propio negocio pertenece por definición al sector privado.	668	Antes: 7 663 filas x 30 col. Después: 7 663 filas x 30 col.	trabajo\$sector[is.na(trabajo\$sector)] <- 'Privado'
15	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	horas_semana_sec, ingreso_laboral_sec	Valor nulo (columna con más de 50% de faltantes)	4. Valores faltantes	Eliminación de 2 columnas: horas_semana_sec (94.7%), ingreso_laboral_sec (95.9%).	Criterio de la metodología: una columna con más de 50% de vacíos no se imputa. Lo rescataba (si la persona tiene o no una segunda ocupación) ya está completo en ocupacion_secundaria.	7 663	Antes: 7 663 filas x 30 col. Después: 7 663 filas x 28 col.	select(-all_of(names(pct_na)[pct_na > 50]))
16	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	ingreso_laboral	Valor nulo	4. Valores faltantes	Imputación avanzada KNN (VIM::kNN, k = 5, distancia de Gower sobre 14 atributos; se asigna la mediana de los 5 vecinos).	El ingreso depende de educación, ocupación, rama, horas y rol: los 5 registros más parecidos estiman mejor que una mediana global. Se usa la mediana de los vecinos porque la distribución es asimétrica.	3	Antes: 7 663 filas x 28 col. Después: 7 663 filas x 28 col.	VIM::kNN(datos, variable = 'ingreso_laboral', k = 5, dist_var = vars_knn, numFun = median)
17	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	horas_semana	Outlier (error de medición confirmado)	5. Valores atípicos	El valor imposible (más de 168 horas en una semana) se anula (NA) para volver a imputarlo; no se elimina la fila.	Regla física: 7 días x 24 h = 168 h. El registro declara una jornada de más de 24 horas diarias: es un error de digitación, no un valor extremo real.	1	Antes: 7 663 filas x 28 col. Después: 7 663 filas x 28 col.	trabajo\$horas_semana[trabajo\$horas_semana > 168] <- NA
18	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	horas_semana	Valor nulo (generado en el paso 5)	4. Valores faltantes (retorno desde el paso 5)	Imputación estadística: mediana de horas_semana de los registros con el mismo rol.	Recursividad de la metodología: anular el error creó un faltante. Se usa la mediana (y no la media) porque la distribución de horas es asimétrica y tiene colas largas.	1	Antes: 7 663 filas x 28 col. Después: 7 663 filas x 28 col.	ifelse(is.na(horas_semana), tapply(horas_semana, rol, median, na.rm = TRUE)[rol], horas_semana)
19	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	horas_semana	Outlier	5. Valores atípicos	Winsorización superior: las jornadas mayores a media + 3 desviaciones (92.3 h) se reemplazan por ese límite.	Se eligió el criterio Z-score porque el IQR marcaba 1 003 registros (jornadas parciales y de 60-70 h que son reales). Jornadas de 24 h diarias durante 6 o 7 días son inverosímiles pero no se pueden corregir: se capan para que no dominen el análisis.	65	Antes: 7 663 filas x 28 col. Después: 7 663 filas x 28 col.	pmin(horas_semana, mean(horas_semana) + 3 * sd(horas_semana))
20	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	ingreso_laboral	Outlier	5. Valores atípicos	Winsorización al percentil 1 y 99 calculados dentro de cada rol (Empleado: 519.6 a 14 037.77 Bs; Empleador: 322.74 a 22 257.93 Bs).	Los ingresos extremos son plausibles (no se eliminan) pero distorsionan medias y distancias. Se capa por rol porque el empleador gana más por naturaleza: un límite común borraría justo la diferencia que se quiere estudiar.	146	Antes: 7 663 filas x 28 col. Después: 7 663 filas x 28 col.	group_by(rol) %>% mutate(ingreso_laboral = pmin(pmax(ingreso_laboral, quantile(ingreso_laboral, .01)), quantile(ingreso_laboral, .99)))
21	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	log_ingreso_laboral	Outlier (distribución asimétrica)	5. Valores atípicos	Transformación logarítmica: nueva columna log_ingreso_laboral = ln(ingreso_laboral). La asimetría baja de 2.01 a -0.32.	El logaritmo comprime la cola derecha y deja una escala casi simétrica, adecuada para algoritmos basados en distancias o que suponen normalidad. Se conserva también el monto en bolivianos.	7 663	Antes: 7 663 filas x 28 col. Después: 7 663 filas x 29 col.	trabajo\$log_ingreso_laboral <- log(trabajo\$ingreso_laboral)
22	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	tam_empresa, tam_empresa_cat	Outlier (distribución asimétrica)	5. Valores atípicos	Discretización: nueva columna tam_empresa_cat (micro, pequeña, mediana, grande). El conteo original se conserva sin capar (máximo 3 000 personas).	IQR marca 793 registros como atípicos, pero una institución con miles de trabajadores es real, no un error. Agrupar en tramos neutraliza el efecto de la cola sin falsear el dato.	7 663	Antes: 7 663 filas x 29 col. Después: 7 663 filas x 30 col.	cut(tam_empresa, breaks = c(0, 9, 19, 49, Inf), labels = c('Micro', 'Pequeña', 'Mediana', 'Grande'))

ID / Paso (#)	Fecha / Hora	Responsable	Dataset	Atributo / Variable	Tipo de Problema	Metodología	Acción Realizada	Criterio / Justificación	Filas Afectadas (N)	Impacto en Dataset	Código / Script
23	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	edad, anios_estudio, antigüedad_anios, dias_semana; horas_dia	Outlier (valor extremo válido)	5. Valores atípicos	Sin tratamiento: los 1 289 valores marcados por IQR en estas 4 variables se conservan. Se elimina horas_dia por redundante.	Trabajadores de 70 años, 35 años de antigüedad o 1 día de trabajo semanal son casos reales que el modelo debe conocer. horas_dia ya está contenida en horas_semana (= dias_semana x horas_dia).	0	Antes: 7 663 filas x 30 col. Después: 7 663 filas x 29 col.	select(-horas_dia)
24	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	tam_empresa	Inconsistencia lógica (regla de negocio)	6. Validación de consistencia	Corrección al mínimo lógico: tam_empresa = 2 en 50 empleados y 3 empleadores que declararon 1 sola persona.	Una relación empleado-empleador necesita por lo menos dos personas. El declarante se contó solo a sí mismo; se capa al mínimo coherente en vez de eliminar la fila.	53	Antes: 7 663 filas x 29 col. Después: 7 663 filas x 29 col.	trabajo\$tam_empresa[trabajo\$tam_empresa < 2] <- 2L
25	2026-10-07 11:30:20	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	rol, sector	Inconsistencia lógica (regla de negocio)	6. Validación de consistencia	Eliminación de filas: empleadores que declaran trabajar en la administración pública.	El Estado no tiene 'patrones o socios': o la categoría ocupacional o el sector están mal registrados y no hay forma de saber cuál. Como rol es la variable objetivo, una etiqueta dudosa contamina el modelo.	3	Antes: 7 663 filas x 29 col. Después: 7 660 filas x 29 col.	trabajo[!(trabajo\$rol == 'Empleador' & trabajo\$sector == 'Público'),]
26	2026-10-07 11:30:21	Saul Mijael Choquehuanca Huanca	empleado_ empleador (subconjunto de trabajo)	Todas	Verificación final	6. Validación de consistencia	Se re-ejecutan las 11 reglas de negocio y el perfilado del paso 1 sobre el resultado: 0 violaciones en la iteración 2, 0 faltantes y 0 duplicados.	El ciclo se repite hasta que una pasada completa no encuentra nada que corregir; recién entonces el dataset se declara limpio.	0	Antes: 7 660 filas x 29 col. Después: 7 660 filas x 29 col.	repeat { v <- validar(trabajo); if (all(v == 0)) break; ... }; perfilar(trabajo)

6 Resultados

6.1 Calidad del dataset antes y después

Tabla 10: Indicadores de calidad del archivo crudo y del dataset limpio.

Indicador	Antes (crudo)	Después (limpio)
Filas	39 497	7 660
Columnas	275	29
Celdas con valor faltante	6 261 409	0
Columnas con algún faltante	229	0
Variables categóricas sin etiqueta	Todas	Ninguna
Jornadas de más de 168 h/semana	1	0
Ingreso laboral máximo (Bs/mes)	54 000	22 258
Asimetría del ingreso laboral	3.30	2.01 (-0.31 en logaritmo)
Violaciones a reglas de negocio	56	0

El dataset limpio tiene 7 660 registros y 29 columnas (Tabla 11): 6 940 empleados y 720 empleadores que, con el factor de expansión, representan a 2 143 477 y 211 359 personas.

Tabla 11: Diccionario del dataset limpio.

Columna	Tipo	Descripción	Origen en el crudo
id_persona	Texto	Clave única de la persona (folio + número de persona)	folio, nro
folio	Texto	Identificador anonimizado del hogar	folio
nro_persona	Entero	Número de orden de la persona dentro del hogar	nro
upm	Texto	Unidad primaria de muestreo	upm
estrato	Texto	Estrato del diseño muestral	estrato
factor_expansion	Decimal	Personas de la población que representa el registro	factor
departamento	Categórica (9)	Departamento de residencia	depto
area	Categórica (2)	Área de residencia	area
sexo	Categórica (2)	Sexo	s01a_02
edad	Entero	Edad en años cumplidos	s01a_03
fecha_nacimiento	Fecha	Fecha de nacimiento en formato ISO-8601	s01a_04a, s01a_04b, s01a_04c
estado_civil	Categórica (6)	Estado civil o conyugal	s01a_10
jefe_hogar	Categórica (2)	Es jefe o jefa del hogar	s01a_05
nivel_educativo	Categórica (7)	Nivel educativo alcanzado	niv_ed
anios_estudio	Entero	Años de estudio aprobados	aestudio
rol	Categórica (2)	VARIABLE OBJETIVO: Empleado o Empleador	s04b_12
categoria_ocupacional	Categórica (3)	Categoría ocupacional original	s04b_12
grupo_ocupacional	Categórica (10)	Grupo ocupacional (COB, 1 dígito)	cob_op
actividad_economica	Categórica (20)	Rama de actividad de la unidad económica (CAEB)	caeb_op
sector	Categórica (3)	Sector de la unidad económica	s04b_13
tam_empresa	Entero	Personas que trabajan en la unidad económica	s04b_14
tam_empresa_cat	Categórica (4)	Tamaño de la unidad económica en tramos	derivada de tam_empresa
antiguedad_anios	Decimal	Años trabajando en la unidad económica	s04b_11aa, s04b_11ab
dias_semana	Decimal	Días trabajados por semana	s04b_15
horas_semana	Decimal	Horas trabajadas por semana en la ocupación principal	phrs
ocupacion_secundaria	Categórica (2)	Tiene una segunda ocupación	s04e_25
ingreso_laboral	Decimal	Ingreso laboral mensual de la ocupación principal (Bs)	yprilab

Columna	Tipo	Descripción	Origen en el crudo
log_ingreso_laboral	Decimal	Logaritmo natural del ingreso laboral	derivada de ingreso_laboral
afp	Categorica (3)	Situación frente al sistema de pensiones	s04f_35, s04f_36

6.2 Perfil del empleado y del empleador

Con los datos limpios la diferencia entre ambos roles se ve con claridad (Tabla 12). Las cifras describen los registros de la muestra, sin expandir.

Tabla 12: Perfil de cada rol en el dataset limpio.

Indicador	Empleado	Empleador
Registros	6 940	720
Edad mediana (años)	36	44
Años de estudio (media)	13.4	11.6
Antigüedad mediana (años)	3	10
Horas por semana (media)	43.9	50.7
Personas en la unidad económica (mediana)	10	3
Ingreso laboral mediano (Bs/mes)	3 250	4 330
Ingreso laboral medio (Bs/mes)	3 882	5 361
Mujeres (%)	40.1	28.6
Jefes o jefas de hogar (%)	55.2	76.4
Con educación superior (%)	55.5	37.9
En microempresas de 1 a 9 personas (%)	48.9	95.4
Afiliados que aportan a la AFP (%)	45.4	6.5
Con segunda ocupación (%)	4.9	9.2

- **El empleador es mayor y lleva más tiempo en su negocio:** mediana de 44 años frente a 36, y 10 años de antigüedad frente a 3.
- **Dirige unidades muy pequeñas:** el 95.4 % de los empleadores está en microempresas, frente al 48.9 % de los empleados (Figura 8).
- **Gana más, pero con más dispersión:** ingreso mediano de 4 330 Bs frente a 3 250 Bs, con una cola alta más pesada (Figura 9).
- **Está menos protegido:** solo el 6.5 % aporta a la AFP, frente al 45.4 % de los empleados (Figura 10).
- **Tiene menos estudios formales:** 37.9 % con educación superior frente a 55.5 % (Figura 11). El empleado típico es un asalariado calificado de una institución mediana o grande; el empleador típico, el dueño de un negocio pequeño.

Tamaño de la unidad económica por rol

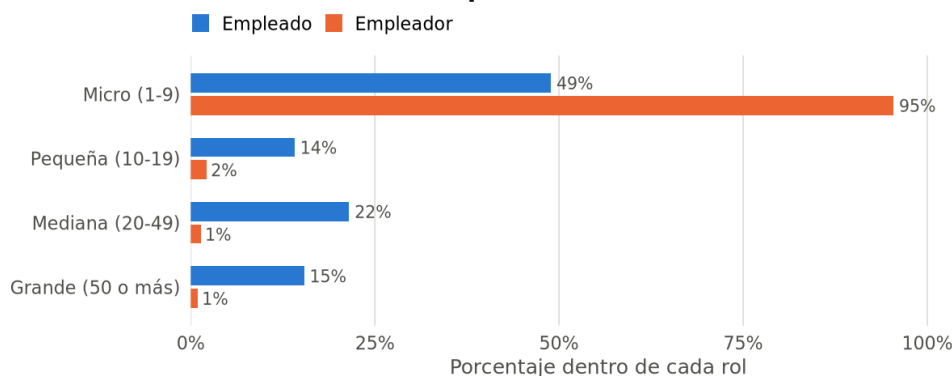


Figura 8: Tamaño de la unidad económica por rol.

Distribución del ingreso laboral por rol

Porcentaje de cada rol por tramo de 1 000 Bs; el último punto agrupa 16 000 Bs o más

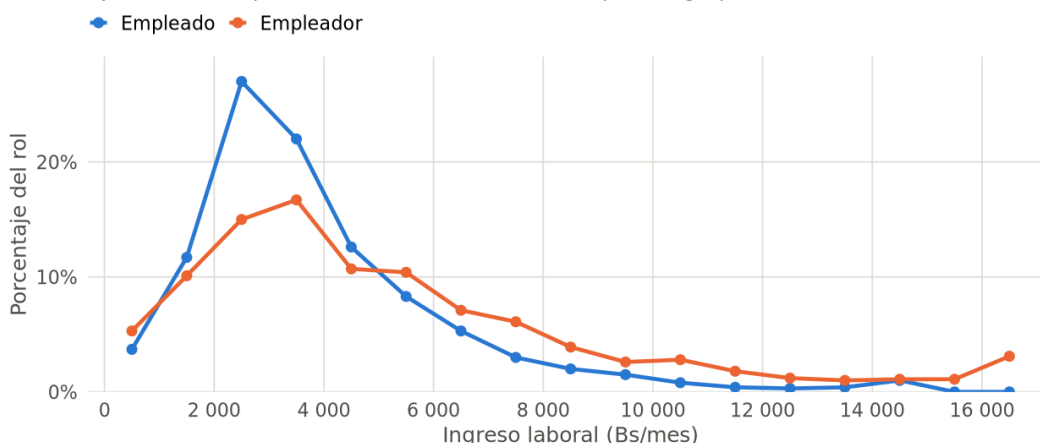


Figura 9: Distribución del ingreso laboral por rol.

Situación frente a la AFP por rol

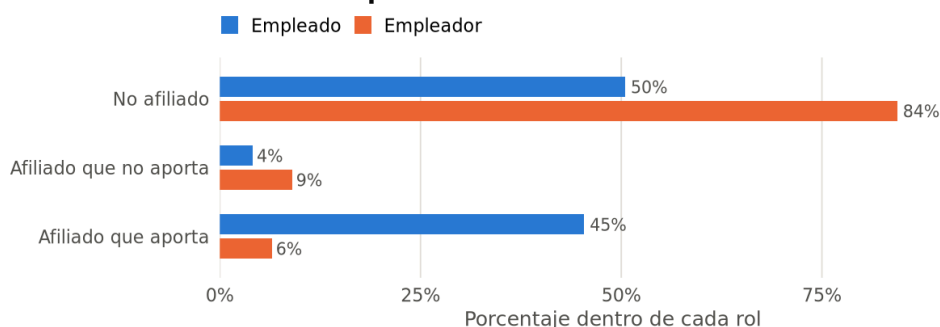


Figura 10: Situación frente a la AFP por rol.

Nivel educativo por rol

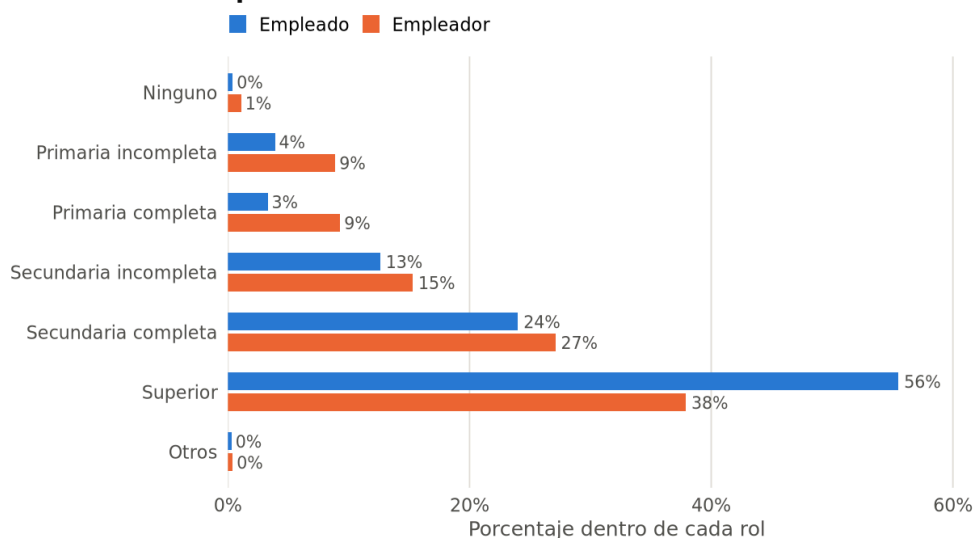


Figura 11: Nivel educativo por rol.

6.3 Relevancia de los atributos

Para comprobar que el dataset sirve al objetivo se midió la asociación de cada atributo con rol (Figura 12): V de Cramér para los categóricos y eta (razón de correlación) para los numéricos. Los más informativos son grupo_ocupacional, tam_empresa_cat, actividad_economica, afp y antigüedad_años. El departamento y el área casi no discriminan. La asociación de sector debe

leerse con cuidado, porque en parte la produce la limpieza: tras la regla R08 ningún empleador queda en el sector público.

Fuerza de asociación de cada atributo con el rol (0 = ninguna, 1 = total)

V de Cramér para atributos categóricos y eta para numéricos

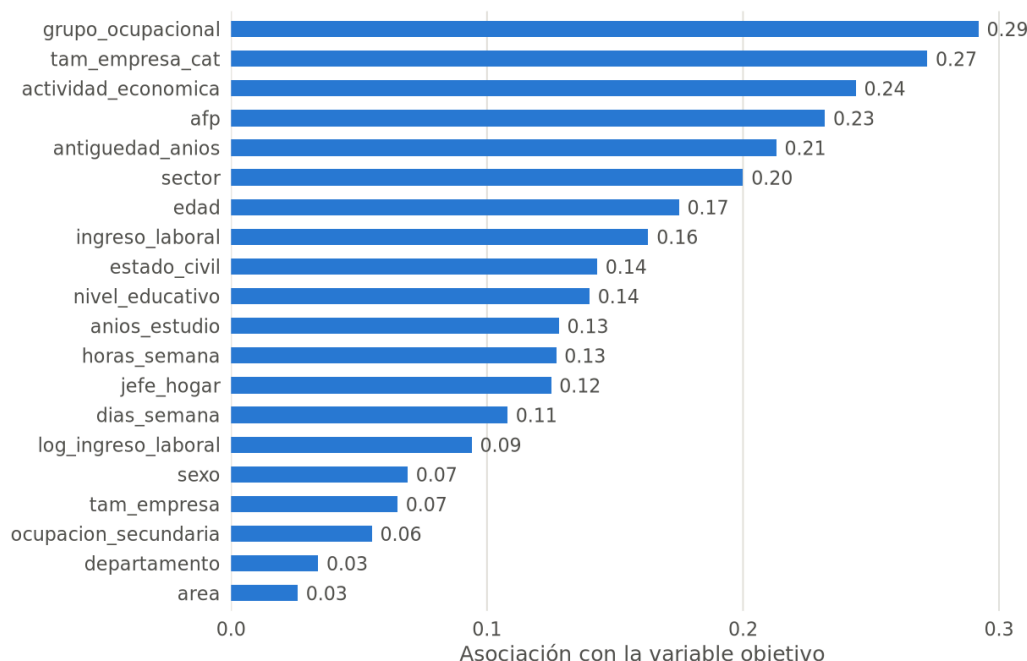


Figura 12: Fuerza de asociación de cada atributo con la variable objetivo.

7 Conclusiones

1. **Se cumplió el objetivo.** De 39 497 registros y 275 columnas se obtuvo un dataset de 7 660 × 29, sin faltantes, sin duplicados y sin violaciones a las 11 reglas de negocio, con 26 cambios documentados en la bitácora.
2. **La limpieza depende del objetivo.** La mayor reducción no se debe a errores sino a pertinencia: 31 823 filas y 242 columnas quedan fuera porque no pertenecen a la relación empleado–empleador. Con otro objetivo la misma base se limpiaría de otra manera.
3. **No todo vacío es un dato perdido.** La mayoría de los faltantes eran estructurales (saltos del cuestionario) y se resolvieron con reglas lógicas o con una categoría explícita; solo 3 valores necesitaron una imputación estadística.
4. **No todo atípico es un error.** Solo un valor era imposible. El resto se capó, se transformó o se conservó según lo que representaba; eliminar todo lo que marca el IQR habría borrado más de mil jornadas reales.
5. **Las reglas del dominio encuentran lo que la estadística no ve.** Empleadores en el sector público y empresas de una sola persona son valores normales para cualquier detector de atípicos; solo la lógica del negocio los delata.
6. **Para la etapa de modelado** quedan dos advertencias: las clases están desbalanceadas (9.6 empleados por cada empleador), por lo que conviene estratificar o ponderar; y las estimaciones de población deben usar `factor_expansion`.

8 Entregables

- **Informe** (este documento) con la bitácora completa en la Sección 5.

- **Código R:** 01_limpieza.R y 02_resultados.R. Para reproducir: colocar persona.csv en la misma carpeta y ejecutar Rscript 01_limpieza.R y luego Rscript 02_resultados.R.
- **Dataset limpio:** empleado_empleador_limpio.csv, y la bitácora en bitacora_limpieza.csv.
- **Tablero** con la explicación, la bitácora filtrable y un explorador del dataset limpio: <https://saulchoque.github.io/mineria-datos-empleado-empleador/>.
- **Repositorio:** <https://github.com/SaulChoque/mineria-datos-empleado-empleador>.

Referencias

- [1] Instituto Nacional de Estadística (2025). *Encuesta de Hogares 2024 (EH 2024): documentación y diccionario de variables*. Archivo Nacional de Datos, catálogo 163. <https://anda.ine.gob.bo/index.php/catalog/163>
- [2] *Metodología para el pre procesamiento de datos y Bitácora de Limpieza de Datos*. Material de la asignatura INF 311, Universidad Mayor de San Andrés, 2026.
- [3] Kowarik, A. y Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, 74(7), 1–16.
- [4] Han, J., Kamber, M. y Pei, J. (2012). *Data Mining: Concepts and Techniques* (3.^a ed.), capítulo 3: Data Preprocessing. Morgan Kaufmann.
- [5] R Core Team (2026). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Viena.