Problem Set 1

Big Data & Machine Learning

Profesor: Ignacio Sarmiento

Grupo 9
Santiago Melo - 202324572
Miguel Blanco - 202412541
María Bernal - 1105793357
Diana Lopera - 1130639521

1. Introduction

Este estudio emplea información obtenida de la Gran Encuesta Integrada de Hogares (GEIH) de 2018, recolectada por el Departamento Administrativo Nacional de Estadística (DANE). Si bien la GEIH proporciona datos a nivel nacional, abarcando tanto áreas urbanas como rurales, y permite el empalme de información sobre ingresos, gasto y otros indicadores relevantes para la medición de la pobreza y la desigualdad, en este ejercicio académico solo se utilizarán los datos para Bogotá del "Reporte de Medición de Pobreza Monetaria y Desigualdad" de 2018.

El objetivo principal del taller es construir un modelo de salarios individuales por hora. Además, se busca analizar la relación entre la edad y el salario, así como la brecha salarial de género. Para evaluar el desempeño del modelo de predicción, se emplearán técnicas de validación, como la validación cruzada.

Los modelos de predicción de ingresos tienen aplicaciones clave en distintos ámbitos. Por un lado, pueden ser una herramienta útil para detectar posibles casos de evasión fiscal mediante la identificación de discrepancias en los ingresos reportados. Por otro lado, estos modelos pueden apoyar a los gobiernos y programas sociales en la identificación de individuos y familias en situación de vulnerabilidad, facilitando la focalización de políticas de asistencia y redistribución de recursos.

Se utilizó web scraping, apoyado en ChromoteSession, para extraer la información de manera eficiente, verificando que las tablas estuviesen completamente cargadas antes de su descarga. Posteriormente, se filtraron los registros para incluir únicamente individuos ocupados mayores de 18 años y se aplicaron técnicas de manejo de datos faltantes y outliers, tales como la imputación de la moda y la mediana, así como la aplicación de winsorizado en la cola inferior del ingreso por hora.

Los resultados descriptivos muestran la alta dispersión en los ingresos laborales por hora, lo que indica una desigualdad salarial marcada. A pesar de que el ingreso laboral por hora promedio es de 7.383 COP, la diferencia notable entre la mediana (5.055,56 COP) y el valor máximo (350.583,3 COP) subraya la presencia de valores atípicos y una distribución desigual de los ingresos.

Además, se observa una clara asociación entre el nivel educativo, el estrato socioeconómico, el tamaño de la empresa y los ingresos promedio. A mayor nivel educativo, estrato socioeconómico y tamaño de la empresa, mayores son los ingresos promedio. Los trabajadores de estratos bajos (1 y 2) tienen ingresos promedio inferiores a 5.500 COP, mientras que en los estratos más altos (5 y 6) superan los 16.000 COP. Quienes no han cursado educación formal o solo tienen primaria perciben ingresos inferiores a 5.000 COP, mientras

que los trabajadores con educación terciaria duplican este monto, alcanzando 10.638 COP por hora. Los empleados en grandes compañías ganan en promedio más del doble que los trabajadores independientes o de microempresa.

Se construyeron modelos de regresión para analizar el perfil edad-salario, revelando una relación cuadrática en la que el salario aumenta con la edad, aunque a una tasa decreciente. En efecto, se observó que un aumento de un año en la edad implica un incremento del 3,8 % en el salario; sin embargo, este efecto es marginalmente decreciente en 0,4 % por cada año que pasa.

Adicionalmente, se exploró la brecha salarial de género en sus formas condicional e incondicional, xxxxxxxxxxxxx. Finalmente, el modelo de predicción de salario que incorpora progresivamente variables explicativas – incluyendo interacciones y transformaciones polinómicas – mejora de forma significativa el rendimiento predictivo de los modelos de ingresos. En concreto, el Modelo 8, que integra factores como edad, nivel educativo, ocupación, estrato socioeconómico y otras interacciones, presenta el menor error de predicción (RMSE = 0,44002), lo que evidencia que una especificación más completa captura mejor la estructura subyacente de los datos.

2. Datos

Este estudio emplea información obtenida de la Gran Encuesta Integrada de Hogares (GEIH) de 2018, solo se utilizarán los datos para Bogotá del Reporte de Medición de Pobreza Monetaria y Desigualdad"de 2018.

El conjunto de datos está conformado por un total de 32,177 registros, divididos en 10 fragmentos y 178 variables, tanto originales como construidas. Estas variables están principalmente relacionadas con información sociodemográfica y laboral de los individuos, incluyendo edad, género, educación, tipo de empleo (formal o informal), horas trabajadas, ingresos salariales y no salariales, cotización a seguridad social y beneficios como subsidios y pensiones, además de información de ocupación.

Para la descarga de la información, fue necesario emplear web scraping [1]. El principal reto en este proceso fue garantizar que las tablas de cada una de las 10 páginas estuvieran completamente cargadas antes de su extracción. Para asegurar una correcta extracción de los datos, se utilizó ChromoteSession para navegar por cada una de las páginas y verificar que la tabla hubiera cargado antes de intentar extraerla. Se implementó una espera activa (*while loop*) que monitoreaba la presencia de una tabla en el DOM, con un tiempo máximo de espera de 90 segundos (*max_wait*), evitando así que el código intentara leer una tabla que aún no había aparecido en la página. Una vez detectada la tabla, su contenido HTML se extraía y convertía en un objeto *rvest*. Cada tabla extraída se guardó en un archivo CSV individual y, finalmente, todas las tablas se combinaron en un único conjunto de datos.

2.1. Preparación de los Datos

El objetivo principal del taller es construir un modelo de salarios individuales por hora. Además, se busca analizar la relación entre la edad y el salario, la brecha salarial de género. Para evaluar el desempeño del modelo de predicción, se emplearán técnicas de validación como la validación cruzada.

Los modelos de predicción de ingresos tienen aplicaciones clave en distintos ámbitos. Por un lado, pueden ser una herramienta útil para detectar posibles casos de evasión fiscal mediante la identificación de discrepancias en los ingresos reportados. Por otro lado, estos modelos pueden apoyar a los gobiernos y programas sociales en la identificación de individuos y familias en situación de vulnerabilidad, facilitando la focalización de políticas de asistencia y redistribución de recursos.

Para la implementación del modelo, se realizó un proceso de preparación de los datos de la siguiente manera:

- El conjunto de datos fue filtrado para incluir únicamente individuos ocupados mayores de 18 años, quedando un total de 16,542 registros.
- Se exploró la estructura interna de la base, para identificar sus componentes y tipos de datos. Adicionalmente, se emplearon los diccionarios disponibles para entender el contenido de las variables.
- Se identificaron variables potenciales de interés.
- Se identificaron missing values y outliers.
- Se genera una base "limpia" para implementar los análisis.

Identificación de variables relevantes para el análisis

En particular, la identificación de potenciales variables de interés se fundamentó en la teoría económica y en la evidencia empírica. De acuerdo con la literatura sobre determinantes salariales y desigualdad de ingresos, las características individuales, el capital humano y las condiciones del mercado laboral juegan un papel crucial en la explicación de los ingresos laborales (Badel y Peña, 2010; Fernández, 2006; Mincer, 1974; Mincer, 1958; Sabogal, 2012).

[1] https://ignaciomsarmiento.github.io/GEIH2018_sample/

En este sentido, la teoría del capital humano resalta la importancia de la inversión en educación y formación como mecanismos para aumentar la productividad y, en consecuencia, los ingresos (Mincer, 1974). La educación se concibe como una inversión que genera retornos en el mercado laboral, por lo que tanto la educación formal como la informal tienen un impacto positivo en los salarios. Además, una de las variables clave en las ecuaciones salariales basadas en el modelo de Mincer es la experiencia laboral, dado que los ingresos tienden a aumentar con la experiencia, aunque a una tasa decreciente, lo que justifica la inclusión de una relación cuadrática entre edad y salario (Badel y Peña, 2010).

Por otro lado, diversos estudios han documentado la persistencia de brechas salariales de género, atribuibles a factores como diferencias en capital humano, discriminación en el mercado de trabajo y segregación ocupacional. En este contexto, la inclusión de variables como sexo, número de hijos, jefatura de hogar y otros aspectos socioeconómicos resulta fundamental para analizar la brecha salarial entre hombres y mujeres (Badel y Peña, 2010; Sabogal, 2012).

Asimismo, las condiciones laborales y la estructura del mercado de trabajo impactan directamente los niveles salariales. Aspectos como la formalidad del empleo y el tamaño de la empresa determinan el acceso a beneficios laborales y la estabilidad en los ingresos, factores estrechamente relacionados con mayores niveles de bienestar económico (Fernández, 2006). La evidencia empírica indica que los trabajadores en empleos formales y en empresas de mayor tamaño suelen recibir salarios más altos y gozar de mejores condiciones laborales, lo que justifica la inclusión de estas variables en un modelo de predicción de ingresos (Aleán-Romero, 2022).

En este orden de ideas, si bien este ejercicio no pretende establecer relaciones causales, sí resulta clave en la identificación de variables relevantes que pueden influir en la predicción del ingreso salarial. Por ello, la selección inicial de variables a explorar se basó en la teoría económica y en la evidencia empírica.

Variable	**Descripción**
y_ingLab_m_ha	Ingreso laboral salarial (nominal por hora)
log_s2	Logaritmo del salario
sex	Sexo del individuo (1 = Hombre, 0 = Mujer)
age	Edad del individuo
maxEducLevel	Nivel máximo de educación alcanzado
nmenores	Número total de menores (a 6 años) en el hogar
H_Head	Jefe del hogar
estrato1	Estrato según recibo de energía
oficio	Ocupación
relab	Tipo de ocupación
formal	Tipo de empleo (1 = Formal, 0 = Informal)
sizeFirm	Tamaño de la empresa por categorías
cotPension	Cotiza en un fondo de pensión

Cuadro 1: Descripción de Variables

Manejo de datos: missing values y outliers

La recolección de datos primarios, como en el caso de la GEIH, enfrenta desafíos como la falta de respuesta total o parcial, lo que genera datos faltantes (*missing values*). Esto puede deberse principalmente al rechazo a responder ciertas preguntas o a la ausencia de algunos miembros del hogar en el momento de la encuesta.

Asimismo, es posible encontrar respuestas con valores incoherentes, errores en el ingreso de la información por parte de los encuestados, o imprecisiones en la medición, lo que puede dar lugar a valores atípicos. Por

esta razón, antes de estimar los modelos, es fundamental realizar un proceso de preparación y depuración de los datos.

Para este ejercicio, se lleva a cabo una exploración de las variables seleccionadas utilizando el paquete *summarytools*, con el fin de obtener una visión general de las variables. En este análisis, se identifican valores faltantes en *maxEducLevel* y *y_ingLab_m_ha*.

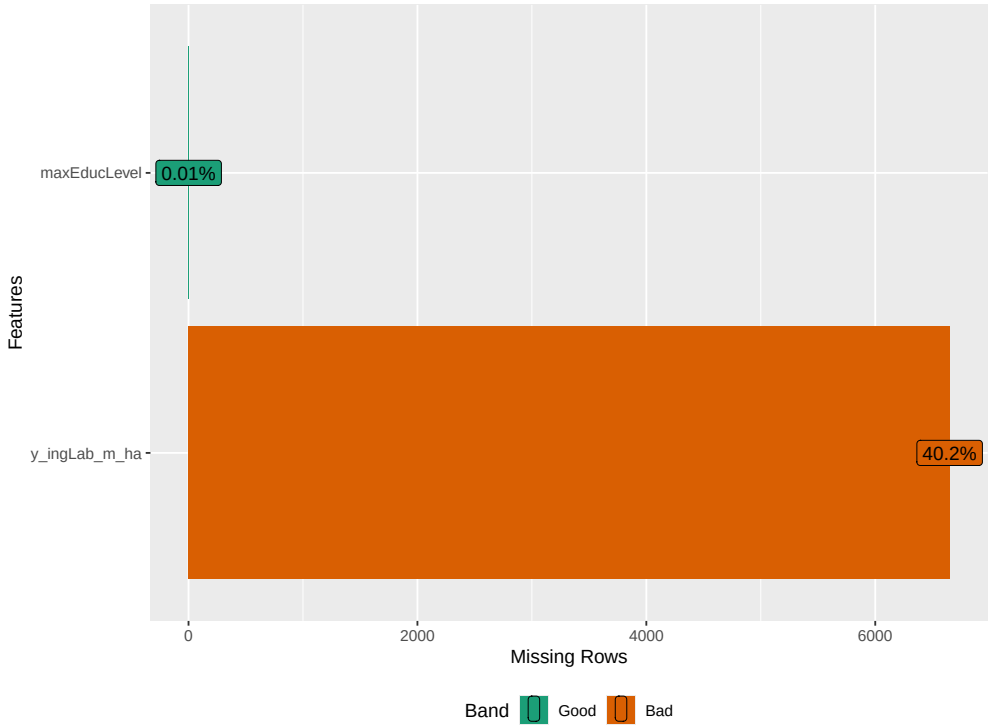


Figura 1: Gráfico de valores faltantes

Para la variable nivel educativo (*maxEducLevel*), al ser categórica, se opta por imputar el valor más frecuente (moda).

En el caso de la variable (*cotPension*), se identificaron 380 observaciones correspondientes a personas pensionadas. Dado que la variable de resultado seleccionada (*y_ingLab_m_ha*[2]) considera todas las ocupaciones, es posible que una persona catalogada como pensionada en una de sus actividades continúe generando ingresos por otra fuente laboral.

Por esta razón, la condición de pensionado no implica necesariamente la ausencia de ingresos laborales adicionales. No obstante, se identificaron 72 registros en los que las personas aparecen como pensionadas pero no tienen información sobre su edad. En estos casos, se recodifican como 1 (cotiza pensión), ya que esto podría deberse a un error en la selección de la variable. Las 308 observaciones restantes podrían corresponder a pensionados "legítimos".

[2]Suma los salarios percibidos por todas las ocupaciones, incluyendo propinas, comisiones, horas extras, primas y bonificaciones, y luego se divide por el total de horas trabajadas. Por ejemplo, si una persona tiene dos trabajos, se suman los salarios de ambos.

	Información de ingreso		
	Con información	Sin información	Total
No cumple edad de pensión	19	53	72
Cumple edad de pensión	88	220	308

Cuadro 2: Frecuencias categoría cotPension (=3 pensionado) e información de ingreso

De acuerdo con la información de la variable tipo de ocupación (*relab*), se identificaron 207 observaciones en la categoría "Trabajador familiar sin remuneración" y 41 observaciones en "Trabajador sin remuneración en empresas o negocios de otros hogares". Ambas categorías corresponden a personas que participan en actividades económicas sin recibir un salario a cambio. Por esta razón, se considera adecuado excluir estos casos del análisis.

Finalmente, en la variable de ingreso salarial, se observa un alto número de valores faltantes (Figura 2), así como una distribución asimétrica con una cola larga a la derecha, lo que sugiere la presencia de valores atípicos o extremos.

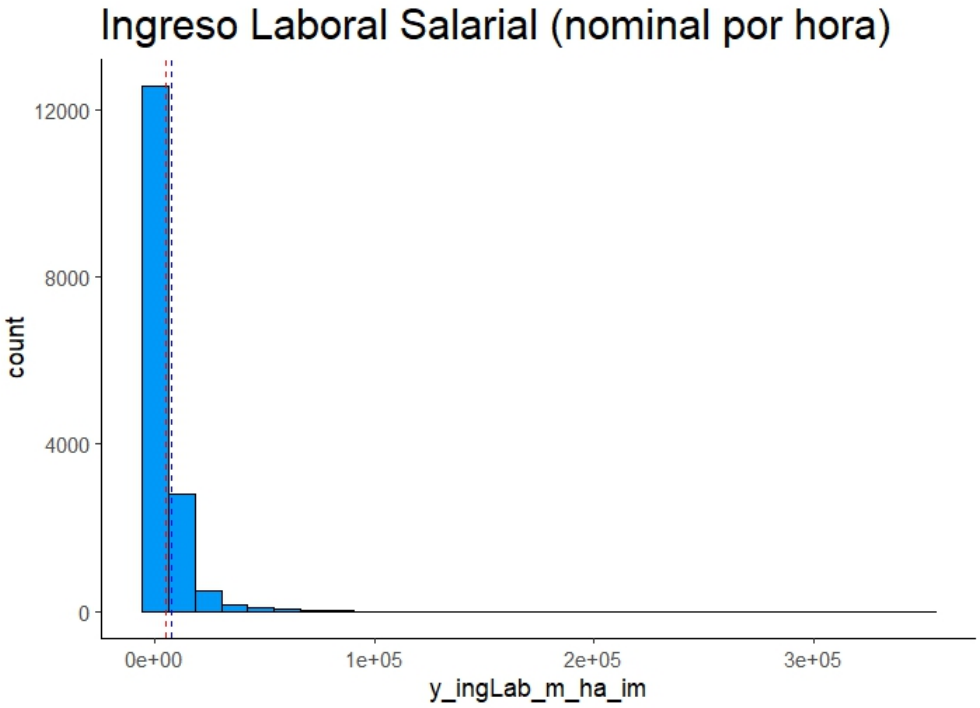


Figura 2: Gráfico de valores faltantes

Dado que la distribución de la variable ingreso tiene una cola larga a la derecha, es más adecuado usar la mediana como método para imputar observaciones faltantes.

	y_ingLab_m_ha_im (Variable imputada)	y_ingLab_m_ha (Variable original)
Min.	326,7	326,7
1st Qu.	4.666,7	4.226,5
Median	5.055,6	5.055,6
Mean	7.342,3	8.822,2
3rd Qu.	5.718,8	8.049,5
Max.	350.583,3	350.583,3
NA's	-	6.402

Cuadro 3: Resumen estadístico de la variable de ingreso laboral con y sin imputación

Se observa que, tras la imputación utilizando la mediana, el valor central de la distribución del ingreso no se ve alterado, ya que la mediana permanece en 5.055,6 (Cuadro 3).

Si bien la imputación de datos permite abordar el problema de valores faltantes, la variable de resultado, ingreso laboral por hora, continúa mostrando una marcada asimetría. La mediana es de 5.055,6, mientras que el valor máximo alcanza los 350.583,3, lo que sugiere que la mayoría de los ingresos se concentran en rangos bajos, mientras que algunos valores extremadamente altos elevan la media.

A pesar de ello, los valores más elevados observados, como el máximo de 350.583,3, siguen siendo plausibles dentro del contexto económico del país. Sin embargo, los valores mínimos reportados, como 326,7 pesos por hora, resultan cuestionables incluso en un país de bajos ingresos como Colombia.

Según el DANE (2024), la incidencia de pobreza en el país durante 2022-2023 fue del 33 %, y la línea de pobreza monetaria per cápita nacional se situó en 435.375 pesos mensuales. Para poner esto en perspectiva, si una persona percibe este ingreso mensual y trabaja 50 horas a la semana, su ingreso por hora sería aproximadamente 2.176,88 pesos. Esto refuerza la idea de que los valores más bajos reportados en la base de datos podrían deberse a errores de digitación (por ejemplo, la omisión de un cero) o inconsistencias en la recolección de información. No obstante, debido a la falta de información adicional para corroborar estas hipótesis, es difícil determinar con certeza la causa exacta de estas irregularidades.

Ante esta situación, se considera pertinente excluir del análisis los registros con valores extremadamente bajos, ya que podrían distorsionar el modelo de predicción. Para ello, se implementa un winsorizado [3] en la cola inferior del ingreso laboral mensual por hora imputado (*y_ingLab_m_ha_im*). Específicamente, los valores por debajo del percentil 5 son reemplazados por el valor correspondiente a dicho percentil, con el objetivo de suavizar la distribución y reducir el impacto de valores atípicos extremadamente bajos. Este ajuste permite obtener resultados más robustos y representativos de la realidad del mercado laboral.

[3] El winsorizado es una técnica estadística utilizada para reducir el impacto de valores atípicos en una distribución. Consiste en reemplazar los valores extremos (tanto en la cola inferior como en la superior) por un valor más cercano dentro de un umbral predefinido, en lugar de eliminarlos completamente.

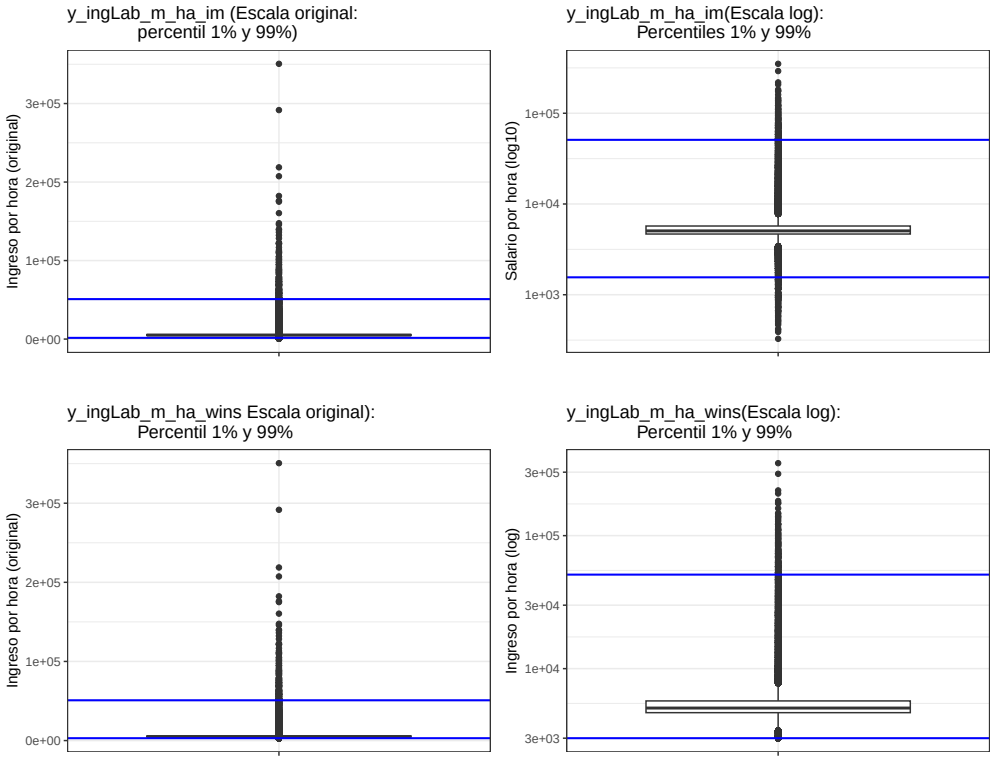


Figura 3: Boxplot: Outliers de la variable de *Ingreso Laboral*

La Figura 3 presenta los diagramas de caja del ingreso por hora (imputado y winsorizado), tanto en su escala original como en la logarítmica. En la escala original, los ingresos exhiben una gran dispersión, con una concentración significativa de observaciones en la parte inferior de la distribución y la presencia de valores atípicos que alcanzan cifras considerablemente altas. Esto sugiere diferencias marcadas en la estructura salarial.

Al aplicar la transformación logarítmica, se obtiene una mejor visualización de la distribución, ya que se reduce la influencia de los valores extremos y se facilita una comparación más clara. Esta transformación permite identificar patrones y tendencias en los datos de manera más efectiva, mitigando la distorsión causada por la asimetría en la distribución del ingreso.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
y_ingLab_m_ha_wins	3000	4667	5056	7383	5719	350583

Cuadro 4: Resumen estadístico de la variable ingreso Laboral winsorizada

Exploración de variables para el modelo de predicción

Para evitar problemas de multicolinealidad, es fundamental seleccionar las variables predictoras que aporten mayor información sin redundancia. Según las pruebas χ^2, las variables relacionadas con el mercado laboral (*relab, formal, sizeFirm, cotPension*) están altamente asociadas, lo que sugiere que algunas de ellas capturan información similar (**Ver Anexo**).

En este contexto, *sizeFirm* y *oficio* resultan más apropiadas que *relab*, *formal* y *cotPension*, ya que podrían reflejar diferencias estructurales en los ingresos laborales y brindar mayor granularidad en la caracterización del empleo. Adicionalmente, la literatura ha confirmado la relación positiva entre el tamaño de la empresa y el salario, pues las empresas más grandes tienden a ofrecer mejores salarios y condiciones laborales, capturando indirectamente la formalidad del empleo (Brown y Medoff, 1989; Criscuolo, 2000). Por otro lado, autores como Oi y Idson (1999) destacan que la diferencia salarial debida al tamaño de la empresa es sustancial. La brecha salarial del 35 % atribuida al tamaño de la empresa es comparable a la brecha salarial de género y mayor que la brecha salarial entre empleados blancos y negros.

Por su parte, la variable *oficio* es un determinante fundamental del salario, ya que refleja diferencias salariales dentro de cada tipo de ocupación, algo que variables como *relab* podrían no captar con la misma precisión. Estudios sugieren que las diferencias salariales están influenciadas no solo por las características de los trabajadores, sino también por los atributos de la industria y la empresa (Dickens y Katz, 1987).

En particular, estudios empíricos realizados para Colombia indican que los determinantes de los salarios y la desigualdad salarial incluyen tanto características sociodemográficas—como la educación, la edad, el género y la presencia de niños menores de 6 años en el hogar—como factores del mercado laboral, entre ellos la distribución de los sectores económicos y los tipos de empleo (Aleán-Romero, 2022; Badel y Peña, 2010). En este sentido, el tipo de trabajo desempeñado por la persona—ya sea en ocupaciones profesionales y técnicas, directivas, administrativas, comerciales, de servicios o en actividades no agrícolas—es uno de los factores que más influyen en la determinación de los salarios por hora (Fernández, 2006).

A continuación, se presenta un análisis descriptivo[4] de las variables relevantes utilizadas en los modelos.

2.2. Análisis descriptivo Datos

De manera general, los resultados descriptivos indican que, en la muestra de individuos seleccionada, la edad promedio es de aproximadamente 39 años, con un rango que va desde los 18 hasta los 94 años.

En términos de composición familiar, el número de menores de 6 años en los hogares varía entre 0 y 4, con un promedio de 0,30 y una mediana de 0. Esto sugiere que al menos el 50 % de los encuestados no tiene niños en el hogar.

La variable de resultado (ingreso laboral por hora) presenta un promedio de 7.383 COP. Sin embargo, como se ha observado previamente, los ingresos muestran una alta dispersión. Esto se confirma con la diferencia sustancial entre la mediana (5.055,56 COP) y el valor máximo registrado (350.583,3 COP), lo que indica la presencia de valores atípicos o una marcada desigualdad salarial en la muestra (*Cuadro* 5).

Variable	**Media**	**Desv. Est.**	**Mín.**	**Mediana**	**Máx.**
Edad (años)	39,37	13,41	18	38,00	94,0
Número de menores	0,30	0,57	0	0,00	4,0
Salario por hora (todas las ocupaciones)	7.383,37	10.187,09	3.000	5.055,56	350.583,3
Número de observaciones			16.294		

Cuadro 5: Estadísticas descriptivas de variables seleccionadas

Por otro lado, respecto a las características sociodemográficas de los encuestados, se observa que el 53,4 % de la muestra está conformada por hombres y el 46,6 % por mujeres, lo que sugiere una distribución relati-

[4]Versión en HTML de las descriptivas básicas disponible en https://cgiar-my.sharepoint.com/:u:/g/personal/d_c_lopera_cgiar_org/EZq_b-zDyu1Gmr0MYtiNpncBA0aicAZyPuDhH3xbnQF8HA?e=twIa4j

vamente equitativa en términos de género.

En cuanto al nivel educativo, la mayor parte de la población ha alcanzado la educación terciaria (42,3 %) o ha completado la secundaria (31,9 %). Sin embargo, aún existe un 14,3 % con educación primaria o menor, lo que puede incidir en los niveles salariales.

El 77,9 % de los individuos pertenecen a los estratos 2 y 3, lo que indica que la mayoría de la muestra se encuentra en niveles socioeconómicos medios. Los estratos más altos (5 y 6) representan solo el 4,6 %, lo que resalta la baja presencia de individuos con mayores niveles de ingreso en la muestra. Esto refuerza la observación de que la mayor parte de los ingresos se concentran en los niveles más bajos de la distribución.

En cuanto a las características del mercado laboral representadas en el tamaño de la empresa, se encontró que el 36,5 % de los encuestados trabaja en empresas con más de 50 empleados, mientras que un 24,8 % es trabajador independiente. Un dato relevante es que casi un 40 % de la muestra trabaja en pequeñas y medianas empresas (PYMEs), distribuyéndose en un 19,2 % en empresas de 2 a 5 empleados, un 12,6 % en empresas de 11 a 50 empleados y un 6,9 % en empresas de 6 a 10 empleados (Cuadro 6).

Variable	**Frecuencia**	**Porcentaje**
Sexo		
Hombre	8.699	53,4 %
Mujer	7.595	46,6 %
Nivel educativo		
Ninguno	116	0,7 %
Primaria incompleta (1-4)	741	4,5 %
Primaria completa (5)	1.476	9,1 %
Secundaria incompleta (6-10)	1.858	11,4 %
Secundaria completa (11)	5.203	31,9 %
Terciario	6.900	42,3 %
Estrato socioeconómico		
Estrato 1	1.735	10,6 %
Estrato 2	6.801	41,7 %
Estrato 3	5.895	36,2 %
Estrato 4	1.120	6,9 %
Estrato 5	319	2,0 %
Estrato 6	424	2,6 %
Tamaño de la empresa		
Trabajador independiente	4.040	24,8 %
2-5 trabajadores	3.133	19,2 %
6-10 trabajadores	1.124	6,9 %
11-50 trabajadores	2.048	12,6 %
Más de 50 trabajadores	5.949	36,5 %

Cuadro 6: Distribución de frecuencias por categoría

Respecto a las ocupaciones (**ver Anexo**), se observa una fuerte concentración en el sector servicios y comercio, con una notable participación en actividades como ventas ambulantes (10,1 %), almacenistas y auxiliares administrativos (5,8 %) y transporte (5,1 %). Este patrón sugiere una estructura laboral orientada hacia el

empleo informal y de menor estabilidad, donde predominan oficios que requieren menores niveles de calificación formal.

Asimismo, se evidencia una baja representación de profesionales altamente especializados, como médicos (0,9 %), economistas (0,4 %) y arquitectos/ingenieros (2,0 %). En contraste, la participación en empleos técnicos y manuales, como albañiles (3,8 %) y electricistas (1,1 %), es moderada.

Con el fin de profundizar en las relaciones entre algunas variables de interés, la figura 4 ilustra la relación entre edad y género. Se observa que la mediana de edad se sitúa en torno a los 40 años en ambos grupos, con valores específicos de 38 años para las mujeres y 37 para los hombres. Esto indica que la muestra está compuesta principalmente por personas en plena edad productiva (7).

Sin embargo, la dispersión es notablemente amplia, con un rango intercuartil ligeramente más extendido en los hombres, lo que sugiere una mayor heterogeneidad en la edad de esta población. También se identifican valores atípicos en ambos grupos, lo que refleja la presencia de individuos de edad avanzada en la fuerza laboral, quienes, como se observó anteriormente, podrían estar asociados a actividades informales o de cuenta propia, predominantes en la muestra.

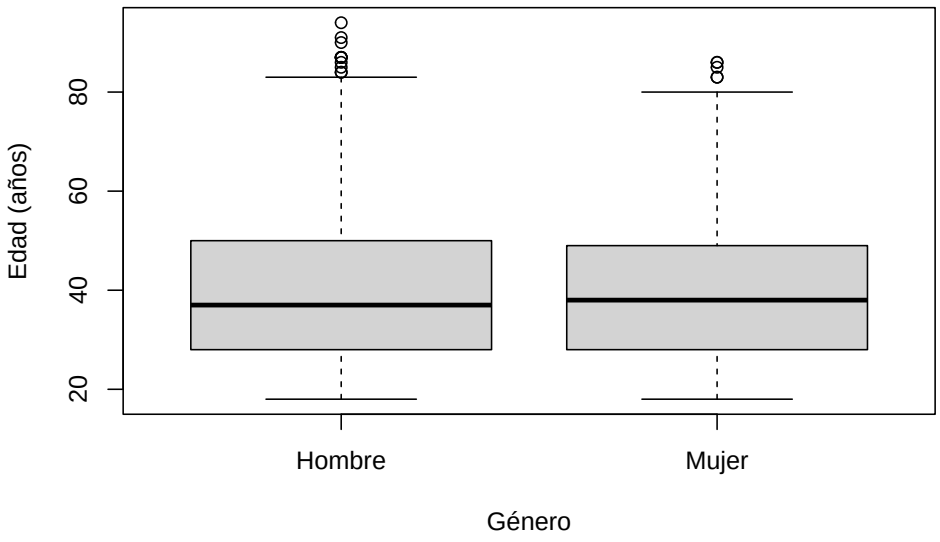


Figura 4: Boxplot edad según género

En cuanto a los ingresos laborales por hora, se observa que las mujeres presentan un ingreso promedio de 7,449.42 COP, solo ligeramente superior al de los hombres (7,325.71 COP). No obstante, el valor de la mediana es igual en ambos grupos (5,055.56 COP), lo que indicaría que la mayoría de la población percibe ingresos por hora en rangos similares. Las diferencias en el promedio son resultado de la asimetría de la variable ingreso, influenciada por valores atípicos (7).

Sexo	Edad media	Edad mediana	Ingreso lab. hora (media)	Ingreso lab. hora (mediana)
Mujer	39,12	38,00	7.449,42	5.055,56
Hombre	39,58	37,00	7.325,71	5.055,56

Cuadro 7: Promedio de ingreso laboral por hora según sexo

Para analizar la distribución del ingreso laboral por hogar según la edad y el género, la figura 5 muestra una tendencia en la que los ingresos aumentan con la edad hasta cierto punto, reflejando el efecto de la acumulación de experiencia. No obstante, a partir de los 50 años, los ingresos tienden a estabilizarse o disminuir.

En cuanto a la brecha de género, no se observa una diferencia sistemática clara en los ingresos entre hombres y mujeres a lo largo de los grupos etarios. Sin embargo, la dispersión del ingreso es mayor en los hombres, especialmente en edades avanzadas, lo que sugiere una mayor presencia de hombres en empleos de alta remuneración.

Por último, los valores extremos en la parte superior del gráfico indican la existencia de individuos con ingresos significativamente elevados, como se ha señalado a lo largo de esta sección, evidenciando una alta heterogeneidad en la distribución salarial.

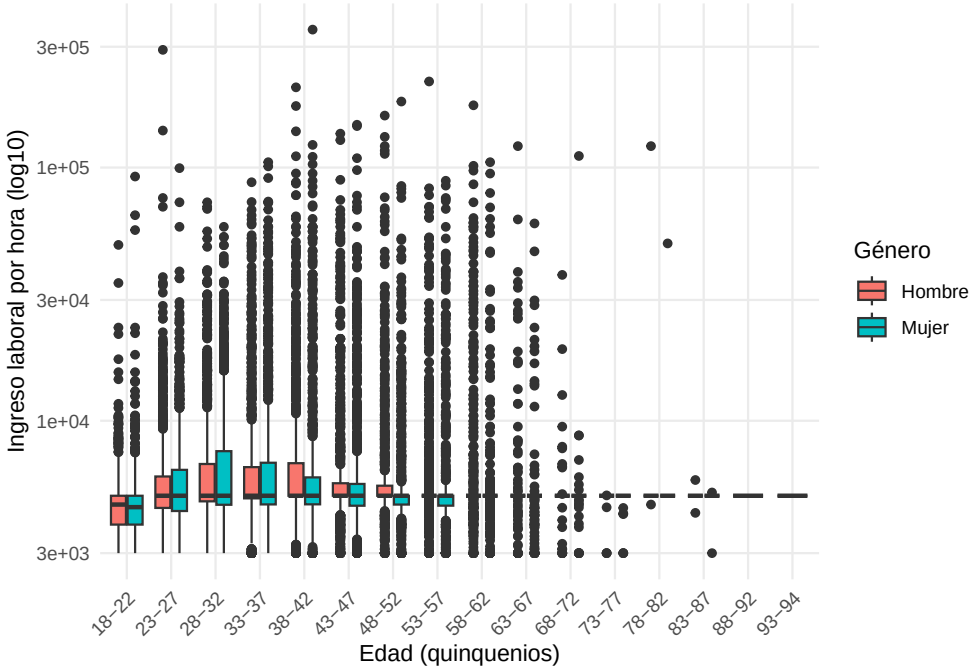


Figura 5: Boxplot: Logaritmo del ingreso laboral por hora: según género y grupos etarios.

Finalmente, el cuadro 8 revela una tendencia clara: a mayor nivel educativo, mayor estrato socioeconómico y mayor tamaño de empresa, mayores son los ingresos promedio.

Los trabajadores de estratos bajos (1 y 2) perciben ingresos promedio inferiores a 5.500 COP, mientras que en los estratos más altos (5 y 6) estos superan los 16.000 COP. De manera similar, el nivel educativo muestra que quienes no han cursado educación formal o solo han completado la primaria perciben ingresos inferiores a 5.000 COP, mientras que los trabajadores con educación terciaria duplican este monto, alcanzando 10.638 COP por hora.

En términos de tamaño de empresa, los empleados en grandes compañías (más de 50 trabajadores) ganan, en promedio, más del doble que los trabajadores independientes o aquellos en microempresas. Esto sugiere que el acceso a empleo formal y estructurado puede desempeñar un papel clave en la determinación de los ingresos.

Categoría	Ingreso promedio (COP)
Estrato socioeconómico	
1	4.983,58
2	5.527,06
3	7.083,60
4	14.455,54
5	16.996,23
6	25.233,07
Nivel educativo	
Ninguno	4.599,16
Primaria incompleta	4.758,42
Primaria completa	4.864,37
Secundaria incompleta	4.844,25
Secundaria completa	5.124,18
Terciario	10.638,21
Tamaño de empresa	
Independiente	4.952,90
2-5 trabajadores	4.910,42
6-10 trabajadores	5.687,93
11-50 trabajadores	7.522,44
Más de 50 trabajadores	10.608,74

Cuadro 8: Promedio de ingreso laboral por hora según estrato socioeconómico, nivel educativo y tamaño de empresa

3. Age-wage profile

En esta sección del taller, llevamos a cabo una regresión que mida la relación entre el salario de las personas, entendido como sus ingresos laborales por hora trabajada, y la edad. Para ello, utilizamos la siguiente ecuación:

$$\log(w_i) = \beta_1 + \beta_2 Age_i + \beta_3 Age_i^2$$

A continuación, se presentan los resultados obtenidos.

- **Resultados de la regresión:**
  Se utilizaron datos de la variable denominada *y_ingLab_m_ha_win*, que corresponde a los ingresos laborales de las personas para cada uno de los trabajos que desempeñan. No se consideran ingresos de trabajo independiente ni otras fuentes de ingresos que pueda tener cada individuo.

  Para esta primera parte del ejercicio, se estimaron dos versiones de la regresión: en la primera, las observaciones con datos faltantes fueron imputadas con la media, mientras que en la segunda, se utilizó la mediana.

Cuadro 9: Logaritmo del salario en función de la edad

	Variable Dependiente:	
	Logaritmo del Salario	
	Media	Mediana
Edad	0,038***	0,033***
	(0,002)	(0,002)
$Edad^2$	−0,0004***	−0,0004***
	(0,00002)	(0,00002)
Constante	7,991***	8,013***
	(0,040)	(0,036)
Observaciones	16.294	16.294
R^2	0,044	0,020
R^2 Ajustado	0,044	0,020
Error Estándar Residual (gl = 16.291)	0,580	0,534
Estadístico F (gl = 2; 16.291)	379,137***	163,275***
Nota:	*p<0,1; **p<0,05; ***p<0,01	

Decidimos usar esta variable en lugar de la variable de ingresos totales, ya que consideramos que esta última podía estar influenciada por otros factores que no necesariamente reflejan el desempeño profesional de las personas, que es lo que, en última instancia, se busca medir. Además, realizamos todo el ejercicio con la variable de ingresos totales y observamos que los resultados eran contrarios a la teoría, lo que sugiere una posible contaminación de la variable.

- **Interpretación de los coeficientes:**
  En la primera regresión, encontramos que un aumento de un año en la edad implica un incremento del 3,8 % en el salario. Sin embargo, este efecto es marginalmente decreciente en 0,4 % por cada año adicional. En consecuencia, el efecto total es que, por cada año que pasa, el salario aumenta en $(100 * (0,038 - 0,004 * Edad))$ %.
  El resultado para el segundo modelo es similar: cada año adicional de edad genera un aumento del 3,3 % en el salario. No obstante, al igual que en el caso anterior, los rendimientos de la edad son marginalmente decrecientes en 0,4 % por cada año adicional. Por lo tanto, el efecto total en este caso es: $(100 * (0,033 - 0,004 * Edad))$ %.

  Todos los coeficientes son estadisticamente significativos al 1 % lo cual implica que existe evidencia de que hay una relación positiva entre la edad y el salario y que el efecto de la edad decrece con el paso de los años, tal como se esperearia en la teoría económica.

- **Ajuste intramuestra del modelo:**
  En esta sección analizamos el ajuste intramuestra de los dos modelos evaluados utilizando diferentes indicadores. El primer indicador que consideramos es el error cuadrático medio (MSE). Para el modelo con la media, obtuvimos un MSE de 0,3366; lo que indica un buen ajuste. Sin embargo, al estimar el modelo con la mediana, se obtuvo un MSE de 0,2848; lo que sugiere que este modelo es preferible. A continuación, se presenta una tabla con los resultados completos.

MSE Media	MSE Mediana
0,3366948	0,2848895

Cuadro 10: Resultados de MSE

  Por otra parte, podemos analizar el R^2 de cada uno de los modelos. El primer modelo tiene un R^2 de 0.044 y el segundo tiene uno de 0.020. Los resultados de los R^2 ajustados son similares. En estos casos se puede evidenciar que el R^2 del modelo de medias es ligeramente mejor pero aun así los resultados son regulares y los valores del R^2 son muy bajos por lo que con este indicador no se puede deducir un buena ajuste de los datos.
  Por otra parte, se pueden analizar los criterios AIC y BIC. Cuyos resultados son los siguientes.

	AIC	BIC
Modelo Medias	28.511,07	28.541,86
Modelo Medianas	25.788,74	25.819,54

Cuadro 11: Valores de AIC y BIC para los modelos

  En base a estos resultados se puede determinar que el modelo de medianas se ajusta mejor a los datos que el modelo de medias. por lo que se opto por usar finalmente el modelo de medianas y no el otro.

- **Estimación de la edad pico y su intervalo vía Bootstrap:**
  En esta sección se busca primero hallar una edad pico para el máximo salario posible. Luego se encuentra un para esta edad máxima con un intervalo del 95 % de confianza. En un primer momento se desarrollo la siguiente gráfica 6 que muestra el salario estimado para cada edad diferente según el modelo usado en el caso anterior.

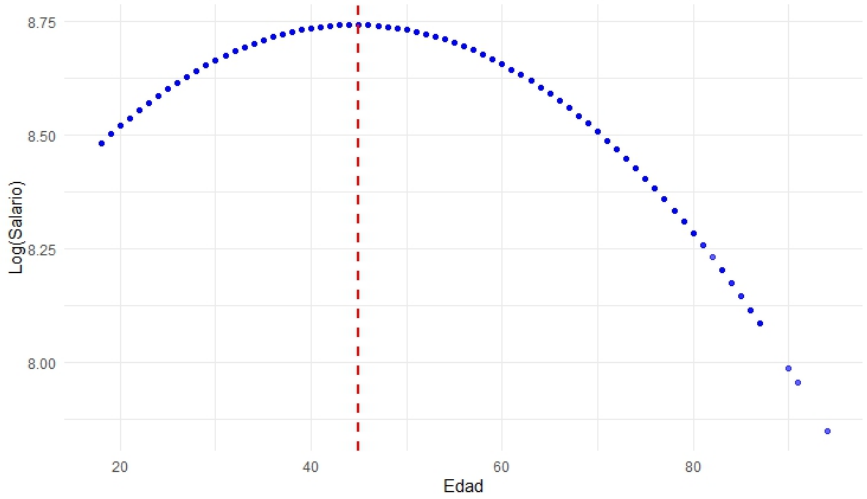


Figura 6: Relación entre el logaritmo de los salarios y la edad estimada por el modelo de medias

Luego se desarrollo un Bootstrap en dos etapas, en una primera etapa se desarrolla una función que estima la edad pico de ingresos y luego se genera un bootstrap de Z=1000 repeticiones. Los resultados fueron de un edad pico estimada en 44.63182, un sesgo de -0.0002658079 y un error estandar de 0.3997789. En la siguiente tabla se presentan los resultados12. Finalmente, se calcularon los intervalos

Estimación	Sesgo	Error Estandar
44.63182	-0.0002658079	0.3997789

Cuadro 12: Resultados de la estimación Bootstrap

de confianza de la edad pico bajo dos metodologías de las cinco disponibles en R. Las metodologías empleadas son Intervalo Percientil, que usa los percentiles de la distribución Bootstrap de la estadística de interés. E Intervalo Corregido y Ajustado por Sesgo, que utiliza una metodología para realizar correcciones por sesgo y por aceleración. En la siguiente tabla 13 se enuncian los dos casos.

Nivel de confianza	Intervalo Percentil	Intervalo Corregido y Ajustado por Sesgo
95 %	(43.91, 45.47)	(43.93, 45.50)

Cuadro 13: Intervalos de confianza

Finalmente, se presenta una gráfica que nos muestra la distribución de los re muestreos del Bootstrap y los intervalos de confianza.

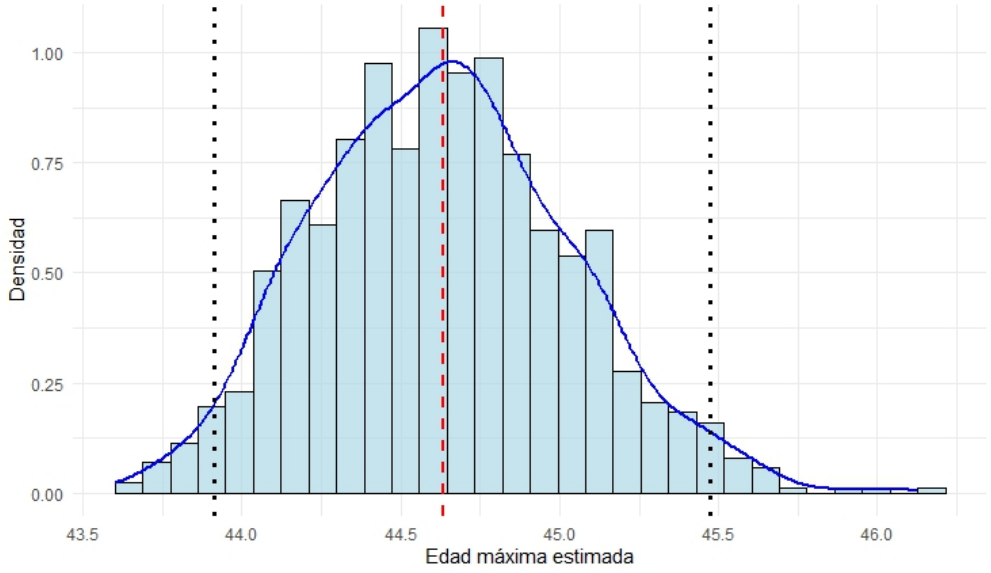


Figura 7: Distribución de Bootstrap con intervalos de confianza calculados por Intervalo Percentil

4. Gender Earning GAP

En esta seccion nos enfocaremos en la brecha salarial de genero.

(a) Estimamos la brecha salarial de genero incondicional con la siguiente ecuacion que relaciona el salario de las personas, entendido como sus ingresos laborales por hora trabajada y el genero:

$$log(w_i) = \beta_1 + \beta_2 Female_i + u_i$$

donde $log(w_i)$ y $Female$ es una variable que toma el valor de 1 cuando el individuo de la muestra es identificado como mujer.

A continuacion se presentan los resultados obtenidos con la brecha salarial de genero incondicional:

Cuadro 14: Logaritmo del salario en función del género

	Variable Dependiente:
	Logaritmo del salario
Mujer	0,002
	(0,008)
Constante	8,666***
	(0,006)
Observaciones	16.294
R^2	0,00000
R^2 Ajustado	−0,0001
Error Estándar Residual	0,539 (gl = 16.292)
Estadístico F	0,035 (gl = 1; 16.292)
Nota:	*p<0,1; **p<0,05; ***p<0,01

Similar al punto anterior, para estimar los ingresos laborales utilizamos la variable $y\_ingLab\_m\_ha\_win$, que corresponde a los ingresos de las personas para cada uno de los trabajos que desempenan. No se consideran ingresos de trabajo independiente ni otras fuentes de ingresos que pueda tener cada individuo.

Para hallar la brecha salarial incondicional la estimacion de los ingresos laborales se hace utilizando unicamente la varibale $Female_i$ que toma el valor de 1 cuando el individuo se identifica como mujer. No obstante, este modelo que depende unicamente de $Female_i$ no arroja resultados estadisticamente significativos, lo que impide interpretar correctamente la relacion cuando se controla por esta unica variable. En particular, en nuestro modelo el hecho de ser mujer no tiene efecto significativo en los ingresos laborales.

En el siguiente apartado, estimaremos la brecha salarial de genero condicional controlando ademas por otras variables.

(b) Dado que en teoria entre empleados con caracteristicas similares y trabajos similares no deberia haber una brecha salarial, estimamos la brecha salarial de genero condicional mediante la siguiente ecuacion:

$$\log(w_i) = \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 Age_i^2 + \beta_5 MaxEduc_i + \beta_6 MaxEduc_i^2 + \beta_7 Ocupacion_i + \beta_8 Estrato_i + \beta_9 Nmenores + u_i$$

Las variables control, ademas del genero, son:

- Age_i y Age_i^2: Utilizamos estas variables para capturar los efectos no lineales de la relacion entre ingreso salarial y la edad, dado que los ingresos suelen aumentar con la edad asumiendo una mayor experiencia laboral.
- $MaxEduc_i$ y $MaxEduc_i^2$: Utilizamos el maximo nivel de educativo obtenido para entender la relacion entre la educacion y el salario, teniendo en cuenta que la relacion no es lineal y no tiene incrementos constantes entre las categorias.
- $Ocupacion_i$: Controlamos por la ocupacion del individuo debido a que diferentes trabajos obtienen distintos niveles de ingresos.
- $Estrato_i$: Controlamos por el estrato socioeconomico, que toma valores entre 1 y 6, dado que el contexto economico de una persona se relaciona con la remuneracion que obtiene por su trabajo, ademas de las oportunidades a las que tiene acceso.
- $Nmenores_i$: Utilizamos el numero de menores en el hogar, debido a que esta variable puede impactar el ingreso en salarial, en la medida en que tiene efecto en la participacion laboral, e incluso en la eleccion de empleo.

I. Utilizando FWL:

II. Utilizando FWL con bootstrap:

Diferencias en los errores estandar:

Comparamos las estimaciones y los errores estandar de ambas aproximaciones:

(c) Graficamos el perfil salario-edad y estimamos los picos de edad implicitos con los intervalos de confianza por genero:

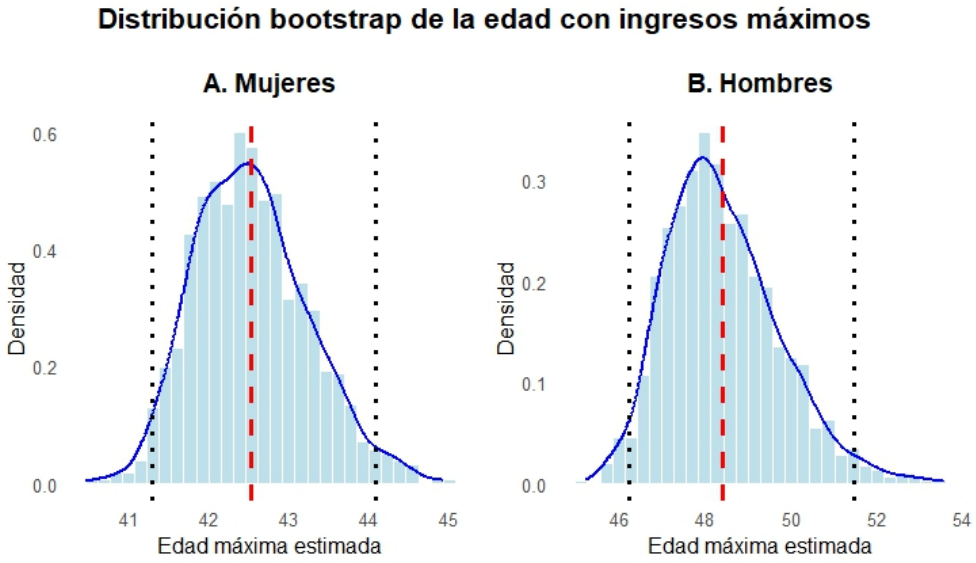


Figura 8: Distribución bootstrap de la edad con ingresos maximos por genero

Conclusiones:

Cuadro 15: Logaritmo del salario en funcion del genero

	Dependent variable:	
	log_s2	y_resid
	(1)	(2)
female	−0.081***	
	(0.007)	
age	0.033***	
	(0.002)	
I(ageˆ2)	−0.0003***	
	(0.00002)	
maxEducLevel_im	−0.271***	
	(0.020)	
I(maxEducLevel_imˆ2)	0.033***	
	(0.002)	
oficio	−0.003***	
	(0.0002)	
relab	−0.067***	
	(0.003)	
estrato1	0.116***	
	(0.004)	
nmenores	0.029***	
	(0.007)	
x1_resid		−0.081***
		(0.007)
Constant	8.352***	0.000
	(0.059)	(0.004)
Observations	16,294	16,294
R^2	0.277	0.007
Adjusted R^2	0.276	0.007
Residual Std. Error	0.459 (df = 16284)	0.458 (df = 16292)
F Statistic	692.722*** (df = 9; 16284)	117.946*** (df = 1; 16292)
Note:		*p<0.1; **p<0.05; ***p<0.01

5. Predicting Earnings

En esta sección, vamos a comparar los errores de predicción de los modelos lineales (cada vez más complejos), utilizando el *Error Cuadrático Medio* (RMSE) como medida de rendimiento predictivo. En este sentido, evaluaremos la precisión de varios modelos de predicción a partir de dos aproximaciones:

- ***Validation Set Approach***: Este método divide el conjunto en dos partes, uno de entrenamiento y uno de validación. El primero se usa para ajustar el modelo, mientras que el segundo se usa para evaluar el rendimiento del mismo.

  Si bien este modelo es fácil de implementar y rápido en términos computacionales, la estimación del error de predicción puede ser muy vribales, ya que depende de una única partición, y es poco eficiente en datasets pequeños, pues no utiliza todos los datos para entrenar.

- ***LOOCV***: Este método es un caso extremo de *k-fold cross-validation*, donde $k = n$. Aquí se entrena el modelo con todas las observaciones menos una (la de validación), y se repite el proceso para cada observación de los datos, calculando el error promedio de todas las iteraciones.

  Si bien este modelo usa toda la información disponible para entrenar el modelo y reduce la variabilidad en la estimación del error de predicción, puede ser computacionalmente costoso, especialmente si la muestra de datos es grande.

Para esto, se compara el rendimiento predictivo de todas las especificaciones anteriores, incluyendo 5 especificaciones adicionales, que introduzcan cambios polinomiales (no linealidades) y aumentos en la complejidad del modelo. Los modelos a comparar son los siguientes:

- **Modelo 1: Age Wage Profile**

$$\log(w_i) = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \varepsilon_i$$

- **Modelo 2: Gender Earning GAP (No controls)**

$$\log(w_i) = \beta_0 + \beta_1 Female_i + \varepsilon_i$$

- **Modelo 3: Gender Earning GAP (Controls)**

$$\begin{aligned}\log(w_i) = \beta_1 + \beta_2 Female_i + \beta_3 Age_i + \beta_4 Age_i^2 + \beta_5 MaxEduc_i + \beta_6 MaxEduc_i^2 \\ + \beta_7 Ocupacion_i + \beta_8 Estrato_i \beta_9 Nmenores + \varepsilon_i\end{aligned}$$

- **Modelo 4:**

$$\log(w_i) = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Female_i + \varepsilon_i$$

- **Modelo 5:**

$$\log(w_i) = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Female_i + \beta_4 (Age_i * Female_i) + \varepsilon_i$$

- **Modelo 6:**

$$\begin{aligned}\log(w_i) = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Female_i + \beta_4 (Age_i * Female_i) \\ + \beta_5 MaxEduc_i + \beta_6 Ocupacion_i + \beta_7 Estrato_i + \varepsilon_i\end{aligned}$$

- **Modelo 7:**

$$\begin{aligned}\log(w_i) = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Female_i + \beta_4 (Age_i * Female_i) \\ + \beta_5 MaxEduc_i + \beta_6 (MaxEduc_i * Age_i) + \beta_7 Ocupacion_i + \beta_8 Estrato_i + \varepsilon_i\end{aligned}$$

- **Modelo 8:**

$$\begin{aligned}\log(w_i) = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Female_i + \beta_4 (Age_i * Female_i) \\ + \beta_5 MaxEduc_i + \beta_6 MaxEduc_i^2 + \beta_7 (MaxEduc_i * Age_i) + \beta_8 Ocupacion_i \\ + \beta_9 Estrato_i + \beta_{10} Nmenores_i + \beta_{11} (Nmenores_i * Age_i) + \beta_{12} TFirma_i + \varepsilon_i\end{aligned}$$

5.1. Validation Set Approach

Dividimos el conjunto muestral en dos subconjuntos, uno de entrenamiento (*training*) y uno de validación o prueba (*testing*). Estos tendrán una distribución de 70 % y 30 % respectivamente. La cantidad de datos en cada subconjunto se puede ver en 9.

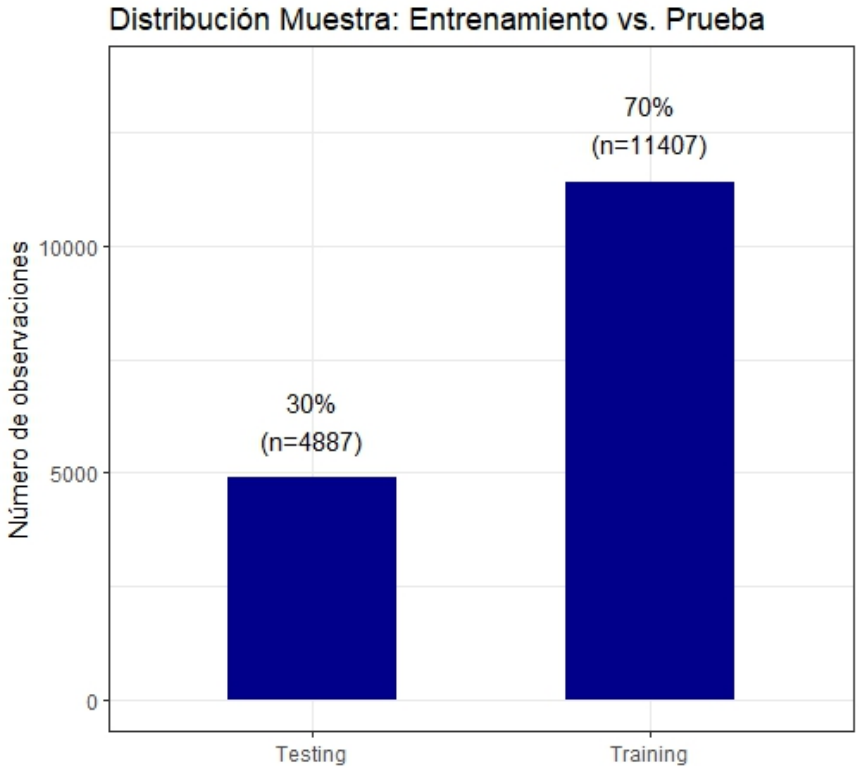


Figura 9: Distribución de la muestra

Se hace la estimación de los modelos a partir del subset de entrenamiento, y obtenemos los siguientes resultados:

Cuadro 16: Variable dependiente: Logaritmo del salario

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Edad	0.031***		0.032***	0.031***	0.033***	0.031***	0.036***	0.028***
	(0.002)		(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
Edad^2	−0.0003***		−0.0003***	−0.0003***	−0.0004***	−0.0003***	−0.0003***	−0.0003***
	(0.00003)		(0.00002)	(0.00003)	(0.00003)	(0.00002)	(0.00002)	(0.00002)
Mujer		0.002	−0.082***	−0.005	0.103***	−0.020	−0.057**	−0.061**
		(0.010)	(0.009)	(0.010)	(0.032)	(0.028)	(0.027)	(0.027)
Constante	8.045***	8.667***	8.403***	8.046***	7.987***	7.546***	7.088***	8.043***
	(0.044)	(0.007)	(0.070)	(0.044)	(0.047)	(0.052)	(0.102)	(0.130)
Controles	Ninguno	Ninguno	Educ, Educ^2, Ocup, T.Ocup	Ninguno	Edad*Mujer	+Educ, Ocup, Estrato	+Educ*Edad	+Educ^2, Niños, Niños*Edad, T.Firma
Observaciones	11,407	11,407	11,407	11,407	11,407	11,407	11,407	11,407
R^2	0.018	0.000	0.275	0.018	0.019	0.232	0.300	0.308
Nota:								*p<0.1; **p<0.05; ***p<0.01

Con respecto al desempeño general de los modelos, este mejora conforme se incorporan más variables explicativas. Dicha mejora se refleja en el incremento del R^2, que parte de un valor de $0,018$ en los modelos más simples y llega a un valor de $0,308$ en el modelo de mayor complejidad, indicando que las variables de este

último modelo logran captar aproximadamente el $30,8\,\%$ de la variabilidad en el logaritmo del salario.

En cuanto al desempeño predictivo a través del RMSE, se obtuvieron los siguientes resultados:

Cuadro 17: Desempeño de predicción

Modelo	RMSE
Modelo 1	0.52646
Modelo 2	0.53283
Modelo 3	0.45216
Modelo 4	0.52645
Modelo 5	0.52627
Modelo 6	0.46576
Modelo 7	0.44202
Modelo 8	**0.44002**

En general, el error cuadrático medio (RMSE) también sigue un patrón de mejora, reduciéndose desde $0,53$ en el modelo 1 hasta $0,44$ en el modelo 8. Esto significa que los modelos más complejos logran reducir la diferencia promedio entre los valores observados y los predichos (note que el modelo 3, debido a la especificación del problema en el punto 4, es más complejo que el modelo 4 y 5), lo que sugiere que capturan mejor la estructura subyacente de los datos.

Sin embargo, se observa que el mayor salto en reducción del RMSE ocurre entre los modelos 5 y 6, y luego entre los modelos 6 y 7. A partir de este punto, la disminución en el RMSE es menos pronunciada. Esto sugiere que las primeras variables agregadas tienen un impacto significativo en la mejora del modelo, mientras que las últimas contribuyen de manera más marginal.

Otro aspecto relevante es que los modelos iniciales contienen únicamente variables generales como la edad y el género. Por una parte, se resalta que aquellos modelos que no contemplan la variable *Edad* y en especial la variable $Edad^2$ tienen un RMSE mayor. Por otra parte, tener pocas variables pueden generar sesgos en la estimación de los coeficientes debido a problemas de variable omitida, por lo que al agregar otras características relevantes del individuo, como el nivel educativo, la ocupación y el estrato socioeconómico, los coeficientes de los regresores iniciales pueden cambiar, reflejando un mejor ajuste del modelo a la realidad económica.

El modelo 8 presenta el menor error de predicción (RMSE = 0.44002), lo que indica que es la especificación más precisa en términos de predicción del logaritmo del salario. La principal diferencia entre este modelo y los anteriores es la inclusión de variables adicionales como el nivel educativo en forma polinomial ($MaxEduc^2$), interacciones con la edad y la presencia de niños en el hogar.

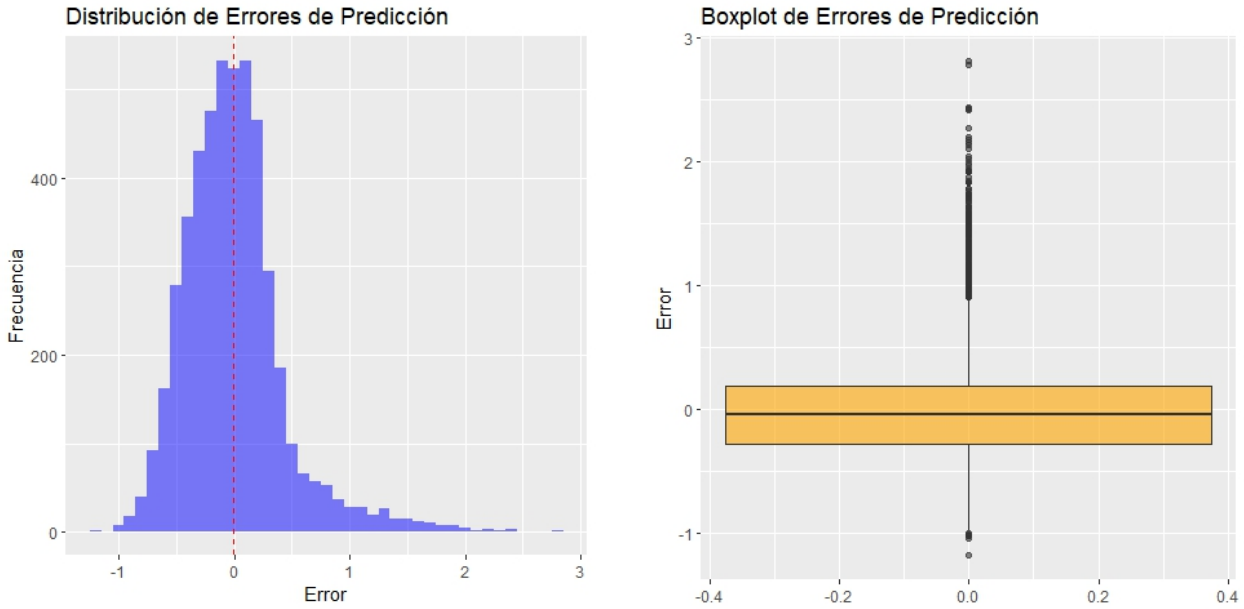
El modelo 7 tiene un desempeño muy cercano al modelo 8 ($RMSE = 0,44202$), lo que sugiere que la incorporación de los últimos controles en el modelo 8 mejora la predicción, pero en menor medida en comparación con las mejoras logradas en modelos anteriores. Es decir, aunque el modelo 8 es el mejor en términos de RMSE, la ganancia marginal respecto al modelo 7 es pequeña.

Desde una perspectiva económica, el modelo 8 parece ser el más completo, ya que incluye factores relevantes que afectan el salario, como el nivel educativo, el tamaño de la firma y la presencia de niños en el hogar. Estos factores tienen sentido dentro de la teoría económica del capital humano y del mercado laboral, lo que respalda su inclusión en la especificación final.

Es posible que el modelo, debido a su complejidad, presente problemas de *overfitting*, el cual ocurre cuando los datos de entrenamiento capturan el ruido en lugar de patrones generalizables. Sin embargo, en este caso no hay evidencia clara de sobreajuste, ya que el RMSE disminuye progresivamente con la inclusión de más variables y no se observa un incremento pronunciado en la varianza de los coeficientes o un deterioro en la capacidad predictiva.

Para confirmar que el modelo no esté sobreajustado, es posible aplicar técnicas como validación cruzada. Si el modelo 8 mantiene un buen desempeño (*performance*) en datos no utilizados en la estimación, es posible pensar que no hay *overfitting*.

Por otro lado, para la especificación con el error de predicción menor, **Modelo 8**, exploramos aquellas observaciones que parecen desviarse de la tendencia general (*miss the mark*), para esto examinamos la distribución de los errores de predicción.



(a) Histograma (b) Boxplot

Figura 10: Distribución de los errores de predicción

De acuerdo con la Figura 10, la distribución del error tiene valores extremos, en el caso del boxplot, estas se encuentran principalmente en la parte superior de esta. Hay varios valores atípicos que superan el rango intercuartil y se extienden lejos de la mediana.

En el histograma, por ejemplo, la mayoría de los errores de predicción están concentrados alrededor de cero, indicando que el modelo tiene un buen *performance* en promedio. Sin embargo, hay una asimetría hacia la derecha, sugiriendo que en algunos casos el modelo está subestimando los valores reales.

Estos valores extremos podrían indicar individuos con ingresos muy altos que el modelo no logra predecir, y que se pueden deber a factores como ingresos no declarados, errores en la información reportada o simplemente características que el modelo no está capturando (variables omitidas).

Si estas observaciones corresponden a ingresos subestimados, podrían representar casos donde el modelo no logra capturar ingresos adicionales no declarados, por ejemplo, de personas con trabajos informales o con múltiples fuente de ingresos, como rentistas de capital. En este sentido, son *outliers* que pueden ser relevantes

para la DIAN.

También es posible que estos valores atípicos correspondan a observaciones con características que el modelo no captura bien, lo que sugiere un problema de especificación del modelo. No obstante, debido a que la cola se encuentra únicamente en la parte derecha de la distribución, es posible pensar que el modelo cuenta con un buen poder predictivo, y que estos datos atípicos corresponden a casos donde no se declaran bien los ingresos, y por ende están subestimados.

Sin embargo, es importante tener en cuenta que los *outliers* podrían ser una mezcla de ambos casos. Para confirmar esto, es necesario examinar más de cerca las características de las observaciones con los errores extremos.

5.2. LOOCV

En esta subsección, procedemos a comparar el desempeño predictivo obtenido bajo la aproximación de *Validation Set Approach* contra los resultados obtenidos de la aproximación de *Leave One Out - Cross Validation*.

Cuadro 18: Comparación de errores de predicción

	RMSE (Validation Set)	RMSE (LOOCV)
Modelo 7	0.44202	0.45027
Modelo 8	**0.44002**	**0.44772**

De acuerdo con la tabla, se puede observar que el Modelo 8 sigue teniendo el menor error de predicción en ambos enfoques, confirmando que es el modelo con mejor ajuste. Sin embargo, los valores de *LOOCV* son un poco mayores que los de *VSA*; esto sucede ya que LOOCV tienden a producir un error mayor debido a su mayor estabilidad y menor varianza al calcular el error promedio de todas las iteraciones. No obstante, el aumento del error bajo LOOCV no es drástico, lo que sugiere que el modelo 8 es relativamente estable.

En este sentido, LOOCV confirma que el modelo 8 es el mejor en términos predictivos, sin embargo, el error de predicción aumenta ligeramente, lo que puede indicar una relación con la presencia de observaciones influyentes, es decir, observaciones que al ser removidas, causan un cambio importante en los coeficientes del modelo.

6. Anexo

6.1. Anexos-Tablas

Cuadro 19: Distribución de datos por tamaño de empresa y formalidad

Tamaño Firma	**0 (No Formal)**	**1 (Formal)**	**Total Filas**
1	3.192	848	4.040
	(0,790)	(0,210)	(0,248)
2	2.269	864	3.133
	(0,724)	(0,276)	(0,192)
3	510	614	1.124
	(0,454)	(0,546)	(0,069)
4	375	1.673	2.048
	(0,183)	(0,817)	(0,126)
5	264	5.685	5.949
	(0,044)	(0,956)	(0,365)
Total Columnas	**6.610**	**9.684**	**16.294**

Chi-Square Test Results:

- Chi-squared: 7.447,33
- Degrees of freedom: 4
- P-value: 0

Cuadro 20: Distribución de datos por pensión y formalidad

Pensión	0 (No Formal)	1 (Formal)	Total Filas
1	2.087 (0,223)	7.255 (0,777)	9.342 (0,573)
2	0 (0,000)	632 (1,000)	632 (0,039)
3	353 (0,611)	225 (0,389)	578 (0,035)
4	3.776 (0,740)	1.330 (0,260)	5.106 (0,313)
5	385 (0,615)	241 (0,385)	626 (0,038)
8	1 (1,000)	0 (0,000)	1 (0,000)
9	8 (0,889)	1 (0,111)	9 (0,001)
Total Columnas	**6.610**	**9.684**	**16.294**

Cuadro 21: Distribución de datos por cotización a pensión y formalidad

Cotización a Pensión	0 (No Formal)	1 (Formal)	Total Filas
1	69 (0,007)	9.385 (0,993)	9.454 (0,580)
2	6.540 (1,000)	0 (0,000)	6.540 (0,401)
3	1 (0,003)	299 (0,997)	300 (0,018)
Total Columnas	**6.610**	**9.684**	**16.294**

Total Observations in Table: 16294

Chi-Square Test Results:

- Chi-squared: 16005.77
- Degrees of freedom: 2
- P-value: 0

Cuadro 22: Distribución de ocupaciones

Variable	**Descripción**	**Frecuencia**	**Porcentaje**
Oficio	Vendedores (ambulantes, a domicilio, entre otros)	1.652	10.10 %
Oficio	Almacenistas, auxiliares administrativos y similares	945	5.80 %
Oficio	Conductores de transporte	831	5.10 %
Oficio	Empleadas domésticas y niñeras	771	4.70 %
Oficio	Directores y gerentes	689	4.20 %
Oficio	Cocineros, camareros y meseros	682	4.20 %
Oficio	Comerciantes y propietarios	661	4.10 %
Oficio	Seguridad (policías, vigilantes, entre otros)	655	4.00 %
Oficio	Albańiles, carpinteros y similares	621	3.80 %
Oficio	Guardianes y porteros	607	3.70 %
Oficio	Docentes	532	3.30 %
Oficio	Técnicos en ingeniería y topógrafos	506	3.10 %
Oficio	Sastres, modistas y similares	501	3.10 %
Oficio	Auxiliares contables y cajeros	481	3.00 %
Oficio	Estibadores y operarios de maquinaria	465	2.90 %
Oficio	Guías de turismo, auxiliares de farmacia y afines	400	2.50 %
Oficio	Ingenieros y arquitectos	319	2.00 %
Oficio	Otros	4.878	29.80 %
Total		**16.294**	**100 %**

Referencias

Aleán-Romero, A. (2022). Los determinantes de la desigualdad del ingreso laboral en cuatro ciudades colombianas: Cartagena, Barranquilla, Bucaramanga y Pereira, 2001-2021. Evidencia de regresiones por cuantiles. *CS*, (SPE1), 117-138.

Badel, A., & Peña, X. (2010). Decomposing the gender wage gap with sample selection adjustment: evidence from Colombia. *Documento CEDE*, (2010-37).

Brown, C., & Medoff, J. (1989). The employer size-wage effect. *Journal of political Economy*, *97*(5), 1027-1059.

Criscuolo, C. (2000). Employer size-wage effect: a critical review and an econometric analysis. *University of Siena Economics Working Paper*, (277).

DANE. (2024). Pobreza Monetaria en Colombia: Principales resultados - Julio 2024 [Consultado el 2 de marzo de 2025]. https://www.dane.gov.co/files/operaciones/PM/pres-PM-2023.pdf

Dickens, W. T., & Katz, L. F. (1987). Inter-industry wage differences and industry characteristics. *Unemployment and the structure of labor markets*, 48-89.

Fernández, M. d. P. (2006). Determinantes del diferencial salarial por género en Colombia, 1997-2003. *Revista Desarrollo y Sociedad*, (58), 165-208.

Mincer, J. (1974). Schooling, experience and earnings Columbia University Press. *New York*.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of political economy*, *66*(4), 281-302.

Oi, W. Y., & Idson, T. L. (1999). Firm size and wages. *Handbook of labor economics*, *3*, 2165-2214.

Sabogal, A. (2012). Brecha salarial entre hombres y mujeres y ciclo económico en Colombia.