Análisis Estadístico del *Stack Overflow Annual Developer Survey 2024*: Selección de Subconjunto Oficial y Estudio del Salario

Mauricio Medina Hernandez
Camilo H. Perez Fleita
Guillermo Hughes Cardona

29 de diciembre de 2025

Resumen

Este informe presenta el tema del proyecto, el dataset crudo seleccionado, las preguntas de investigación y la motivación. Luego documenta la selección del dataset procesado oficial (variables retenidas, excluidas y variables derivadas). Finalmente, se discuten consideraciones metodológicas clave asociadas a valores faltantes y sesgo de reporte del salario, junto con medidas de mitigación.

Índice

1. Tema del proyecto, dataset crudo y preguntas de investigación

Para nuestro proyecto teníamos el interés de encontrar y analizar factores asociados al salario de desarrolladores alrededor de el mundo, haciendo énfasis en la relación de estos con la modalidad de trabajo, además de explorar si existen perfiles tecnológicos (stack) con diferencias salariales claras.

Con esta idea en mente, seleccionamos el *Stack Overflow Annual Developer Survey 2024* como nuestro dataset: la más reciente de las encuestas que realiza la conocida plataforma *Stack Overflow* y que miles de desarrolladores alrededor de el mundo responden cada año.
Archivo: `survey_results_public.csv` **Dimensiones (raw):** $N = 65437$ filas, $P = 114$ columnas. **Enlace oficial del dataset (referencia):**

- https://survey.stackoverflow.co/

Preguntas a responder en la investigación

RQ1: (Comparación de grupos) *¿Difiere el salario anual (en escala logarítmica) entre modalidades de trabajo* `Remote`*,* `Hybrid` *y* `In-person`*?*

RQ2: (Modelo explicativo) *¿Se mantiene la asociación entre* `RemoteWork` *y salario al controlar por país, experiencia, educación y rol?*

RQ3: (Perfiles tecnológicos) *¿Existen perfiles (clusters) de desarrolladores según su stack tecnológico y difieren en salario?*

Variable objetivo principal: `ConvertedCompYearly` (salario anual convertido).
Transformación clave para el análisis: `log_salary` $= \log(1 +$ `ConvertedCompYearly`$)$.

Selección final del dataset procesado

La construcción del dataset procesado oficial se guió por los siguientes criterios:

1. **Relevancia para las RQs:** retener únicamente variables necesarias para responder RQ1–RQ3.
2. **Control de confusión:** incluir covariables con impacto esperado en salario (`Country`, experiencia, rol, educación).
3. **Viabilidad estadística:** evitar variables con pocas respuestas y baja contribución a las RQs.
4. **Interpretabilidad y reproducibilidad:** preferir variables bien documentadas y derivaciones deterministas.

El dataset procesado oficial contiene 20 columnas: 17 variables base seleccionadas + 3 variables derivadas.

Variable	Por qué se retiene y para qué se usa
`ResponseId`	Identificador único para trazabilidad y control de integridad.
`MainBranch`	Define población objetivo (p.ej., rama profesional) y apoya filtros.
`Employment`	Apoya filtros de población (p.ej., empleo formal).
`Country`	Covariable crítica: controla heterogeneidad por país; esencial por sesgo de reporte del salario.
`RemoteWork`	Variable explicativa central para RQ1 y RQ2.

Variable	Por qué se retiene y para qué se usa
`ConvertedCompYearly`	**Variable objetivo** (salario anual).
`YearsCodePro`	Experiencia profesional (predictor clave de salario).
`YearsCode`	Experiencia general programando (contexto/robustez).
`DevType`	Control por rol(es) del desarrollador; útil en RQ2.
`EdLevel`	Control por educación; útil en RQ2.
`Age`	Control descriptivo y posible covariable de robustez.
`OrgSize`	Contexto laboral (tamaño de organización); posible covariable en RQ2.
`LanguageHaveWorkedWith`	Stack (multi-selección); base para perfiles en RQ3.
`DatabaseHaveWorkedWith`	Stack (multi-selección); base para perfiles en RQ3.
`WebframeHaveWorkedWith`	Stack (multi-selección); base para perfiles en RQ3.
`PlatformHaveWorkedWith`	Stack (multi-selección); base para perfiles en RQ3.
`ToolsTechHaveWorkedWith`	Herramientas (multi-selección); base para perfiles en RQ3.

1.0.1. Variables derivadas agregadas

Se agregaron variables derivadas para habilitar análisis y control de calidad:

- `has_salary`: indicador de disponibilidad de salario, definido como

$$\texttt{has\_salary} = \mathbb{I}(\texttt{ConvertedCompYearly} \neq \texttt{NaN}).$$

  Se usa para diagnosticar sesgo de reporte y definir la población efectiva de análisis salarial.

- `YearsCodePro_num`: conversión numérica de `YearsCodePro` para modelado. Para valores no numéricos, se aplica limpieza previa y luego `to_numeric(errors=çoerce")`.

- `log_salary`: transformación del salario anual:

$$\texttt{log\_salary} = \log(1 + \texttt{ConvertedCompYearly}),$$

  utilizada para estabilizar varianza y reducir influencia de colas largas/outliers.

1.1. Variables excluidas: qué se quitó y por qué

Dado que el dataset crudo incluye 114 columnas, se excluyeron variables para reducir ruido, evitar pérdida innecesaria de datos y mantener foco analítico en salario.

1.1.1. Familias de variables eliminadas

- **Bloques de IA** (`AINext*`, `AITool*`, etc.): muy pocas respuestas y no alineadas con RQ1–RQ3.
- **Preferencias/aspiraciones** (`WantToWorkWith`, `Admired`): describen intención, no experiencia real; para RQ3 se prioriza `HaveWorkedWith`.
- **Subdominios específicos con subpoblación reducida** (p.ej., `Embedded*`): incrementan sesgo y reducen generalización para el objetivo salarial.
- **Uso de Stack Overflow / comunidad**: relevantes para otros temas (comportamiento en la plataforma), pero no para el foco salarial.
- **Baterías de hábitos/conocimiento** (`Knowledge_*`, `Frequency_*`): útiles para un proyecto alternativo (satisfacción/hábitos), pero cambian el foco y aumentan dimensionalidad.

2. Consideraciones finales: sesgo, valores faltantes y decisiones de mitigación

2.1. Missingness del salario y población efectiva

La variable objetivo `ConvertedCompYearly` presenta un $\approx 64{,}19\,\%$ de valores faltantes; por construcción, `log_salary` hereda el mismo missingness. En consecuencia, el análisis salarial se interpreta sobre la subpoblación:

"participantes que reportaron compensación anual".

2.2. Evaluación del sesgo de reporte

Se evaluó la asociación entre `has_salary` y variables clave mediante Chi-cuadrado, reportando tamaño de efecto (Cramér's V) para no sobreinterpretar significancia estadística con N grande.

Cuadro 2: Asociación entre `has_salary` y variables explicativas (Chi-cuadrado y Cramér's V)

Contraste	χ^2	p	Cramér's V
RemoteWork vs has_salary	301.69	$3{,}08 \times 10^{-66}$	0.074
Country(top15) vs has_salary	876.18	$5{,}48 \times 10^{-178}$	0.147

Interpretación:

- **RemoteWork**: asociación estadísticamente significativa pero de **tamaño pequeño** ($V \approx 0{,}074$).
- **Country**: asociación **más relevante** (pequeña–moderada; $V \approx 0{,}147$), indicando que el reporte de salario depende del país.

2.3. Implicaciones para el análisis y mitigación

Para mitigar sesgo y mejorar robustez:

1. **Control por país:** `Country` se incluye como covariable en RQ2 y se considera estratificación o restricción a países con suficiente muestra (p.ej., top 15 por frecuencia con salario).
2. **Definición de subconjunto salarial:** análisis de RQ1–RQ2 se realiza con `has_salary=True` y `RemoteWork` no nulo.
3. **Transformación logarítmica:** uso de `log_salary` para reducir la influencia de colas largas/outliers.
4. **Reporte transparente:** las conclusiones se formulan para la subpoblación que reporta salario y se documentan limitaciones.

2.4. Limitaciones

- **Generalización condicionada:** el análisis salarial describe a quienes reportaron salario, no necesariamente a todos los encuestados.
- **Causalidad:** el estudio es observacional; no permite inferencias causales (p.ej., remoto $\Rightarrow$ mayor salario).
- **Alta dimensionalidad en RQ3:** las variables multi-selección requieren codificación (one-hot), lo cual puede incrementar dimensionalidad y requerir regularización/selección.