# Documentación data cleaning Manager Survey

## 1. Variables en la base de datos original

| Nombre de la variable (original) | Tipo de variable | Descripción |
|---|---|---|
| Timestamp | Fecha / datetime | Fecha y hora en la que el encuestado diligenció el formulario. |
| How old are you? | Texto | Edad reportada por la persona encuestada. |
| What industry do you work in? | Texto | Industria o sector económico en el que trabaja la persona. |
| Job title | Texto | Cargo o título del puesto de trabajo actual. |
| If your job title needs additional context, please clarify here: | Texto | Información adicional o aclaración sobre el cargo cuando el título no es suficiente. |
| What is your annual salary? (...) | Número entero | Salario anual bruto reportado por la persona, en la moneda indicada posteriormente. |
| How much additional monetary compensation do you get, if any (...) | Número decimal | Compensaciones adicionales como bonos u horas extra en un año promedio. |
| Please indicate the currency | Texto | Moneda en la que se reportan el salario y las compensaciones. |
| If "Other," please indicate the currency here: | Texto | Detalle de la moneda cuando no se encuentra en las opciones estándar. |
| If your income needs additional context, please provide it here: | Texto | Comentarios adicionales sobre el ingreso reportado. |

| | | |
|---|---|---|
| What country do you work in? | Texto | País en el que trabaja la persona. |
| If you're in the U.S., what state do you work in? | Texto | Estado dentro de EE. UU. donde trabaja la persona, si aplica. |
| What city do you work in? | Texto | Ciudad en la que trabaja la persona. |
| How many years of professional work experience do you have overall? | Texto | Años totales de experiencia laboral profesional. |
| How many years of professional work experience do you have in your field? | Texto | Años de experiencia profesional específicamente en su campo. |
| What is your highest level of education completed? | Texto | Nivel educativo más alto alcanzado por la persona. |
| What is your gender? | Texto | Género con el que se identifica la persona encuestada. |
| What is your race? (Choose all that apply.) | Texto | Raza o etnia reportada; puede contener múltiples valores. |

## 2. Variables luego del data cleaning

| Nombre de la variable (modelada) | Tipo de variable | Descripción |
|---|---|---|
| timestamp | Fecha / datetime | Fecha y hora de registro de la encuesta, heredada de la base original sin modificaciones. |
| age | Texto | Edad reportada por la persona encuestada. Se mantiene como texto debido a la heterogeneidad de formatos en las respuestas. |
| industry | Texto | Industria o sector económico luego de una limpieza básica de texto. |
| job_title | Texto | Cargo o título del puesto de trabajo estandarizado. |
| job_title_context | Texto | Información adicional sobre el cargo cuando el título principal no es suficiente. |

| | | |
|---|---|---|
| annual_salary | Numérica | Salario anual bruto reportado por la persona, expresado en la moneda indicada en la variable currency. |
| bonus_compensation | Numérica | Compensaciones adicionales como bonos u horas extra en un año promedio. |
| currency | Texto | Moneda en la que se reportan el salario y las compensaciones. |
| income_context | Texto | Comentarios adicionales proporcionados por la persona sobre su ingreso. |
| country | Texto | País reportado originalmente por la persona, sin estandarizar. |
| state | Texto | Estado o división administrativa donde trabaja la persona, cuando aplica. |
| city | Texto | Ciudad reportada originalmente por la persona. |
| years_experience_total | Texto | Años totales de experiencia laboral profesional reportados. |
| years_experience_field | Texto | Años de experiencia profesional específicamente en el campo laboral del encuestado. |
| education_level | Texto | Nivel educativo más alto alcanzado por la persona. |
| gender | Texto | Género con el que se identifica la persona encuestada. |
| race | Texto | Raza o etnia reportada; puede contener múltiples valores. |
| country_standardized | Texto | País de trabajo estandarizado a partir de reglas de limpieza y normalización (por ejemplo, variantes de USA unificadas como "United States"). |
| city_clean | Texto | Ciudad de trabajo limpiada eliminando valores inválidos, caracteres especiales y respuestas no geográficas. |

| | | |
|---|---|---|
| salario_anual_cop | Numérica | Salario anual convertido a Pesos Colombianos (COP) usando la tasa de cambio del día del ejercicio. |
| compensaciones_cop | Numérica | Compensaciones adicionales convertidas a Pesos Colombianos (COP). |
| total_compensacion_cop | Numérica | Suma del salario anual y las compensaciones adicionales, expresada en Pesos Colombianos (COP). |

# 3. Paso a paso para actualizar los datos y aplicar el modelado

Esta sección describe el procedimiento que debe seguir una persona para actualizar la base de datos y replicar el modelo diseñado, para esta aplicaciones se utilizó la librería de pandas para limpieza de datos, pero se podrían usar otro tipo de herramientas como excel, power query, SQL, etc.

1. **Descargar el archivo**
   Extraer la base de datos “Ask A Manager Salary Survey 2021” de AskAManager.org **(https://docs.google.com/spreadsheets/d/1IPS5dBSGtwYVbjsfbaMCYIWnOuRmJcbequohNxCyGVw/edit?resourcekey#gid=1625408792)** y cargarlo en la herramienta que se vaya a utilizar para la limpieza.

2. **Renombrar columnas**
   Aplicar el diccionario de renombramiento para simplificar los nombres de las variables y prepararlas para el modelado.

3. **Limpieza de variables de texto**
   a. Limpiar columnas de país y ciudad eliminando espacios, caracteres especiales y respuestas inválidas.
   b. Estandarizar países mediante reglas predefinidas (por ejemplo, todas las variantes de USA → "United States").

4. **Filtrar el archivo**

   Para tener un conjunto de datos más estable se decide utilizar sólo los registros que tengan salarios en USD y que pertenezcan al país United States.

5. **Conversión de variables monetarias**
   a. Definir la tasa de cambio USD → COP correspondiente al día en que se realiza la actualización.
   b. Convertir el salario anual y las compensaciones a COP.

6. **Creación de variables derivadas**
   a. Crear los campos `salario_anual_cop` y `compensaciones_cop`.
   b. Crear el campo `total_compensacion_cop` como la suma de ambos.

7. **Validaciones finales**
   a. Revisar valores nulos.
   b. Verificar rangos de salarios.
   c. Confirmar que las conversiones monetarias sean coherentes.

8. **Guardar la base modelada**
   Exportar el dataset final en un archivo .xlsx para su análisis o visualización

9. **Subir los archivos al repositorio de GitHub**
   a. Verificar que el repositorio local esté correctamente inicializado y conectado al repositorio remoto en GitHub (https://github.com/lucasduenas/viz_storytelling).
   b. Agregar los archivos actualizados (dataset modelado, scripts de limpieza y modelado, y documentación) usando *git add.*
   c. Realizar un commit descriptivo indicando los cambios realizados (por ejemplo: "Actualización de datos y modelado").
   d. Subir los cambios al repositorio remoto utilizando *git push.*
   Este paso garantiza el versionamiento del trabajo, la trazabilidad de los cambios y facilita la colaboración y reproducción del proceso por otros miembros del equipo.