Documentación – Dashboard 2
Modelado de datos

Jesus Alberto Parada Perez

1. Base de datos original

La base de datos "Ask A Manager Salary Survey 2021" es una encuesta realizada por askamanager.org a través de la herramienta Google forms que tiene como objetivo recolectar información voluntaria de los salarios que se pagan en los diferentes puestos de trabajo alrededor del mundo y contribuir, con información anónima, a mitigar esa desinformación. La encuesta, disponible desde el 2021, consulta al público general sobre 17 aspectos como: país, tipo de industria, edad, genero, raza, cargo laboral, salario y bonificaciones, entre otros, y con corte al 7 de febrero de 2023 ha recolectado respuestas de 27.935 personas. A continuación, se presenta una breve descripción:

Nombre de la variable	Tipo	Descripción
Timestamp	Fecha	El cuestionario automáticamente recolecta la hora de realización de la encuesta.
How old are you?	Texto	Rango de edad.
What industry do you work in?	Texto	Sector o industria en la que trabaja.
Job title	Texto	Cargo o nombre del puesto de trabajo.
If your job title needs additional context, please clarify here	Texto	Información adicional del cargo o puesto de trabajo.
What is your annual salary? (You'll indicate the currency in a later question. If you are part-time or hourly, please enter an annualized equivalent -- what you would earn if you worked the job 40 hours a week, 52 weeks a year.)	Número	Salario anual. En caso de laborar medio tiempo o por horas realizar una aproximación al salario de tiempo completo.
How much additional monetary compensation do you get, if any (for example, bonuses or overtime in an average year)? Please only include monetary compensation here, not the value of benefits.	Número	Compensación monetaria anual que recibe y que es adicional al salario.
Please indicate the currency	Texto	Moneda en la que esta expresado el salario y las compensaciones.
If "Other," please indicate the currency here	Texto	Si la moneda no se encuentra en la lista desplegable de la pregunta anterior, se debe especificar la moneda
If your income needs additional context, please provide it here	Texto	Información adicional de los ingresos.
What country do you work in?	Texto	País en el que trabaja.
If you're in the U.S., what state do you work in?	Texto	Si trabaja en USA, seleccionar el Estado en el que desarrolla el trabajo.
What city do you work in?	Texto	Ciudad en la que trabaja.
How many years of professional work experience do you have overall?	Texto	Años de experiencia profesional.

How many years of professional work experience do you have in your field?	Texto	Años de experiencia profesional en el área especifica en la que desarrolla el trabajo.
What is your highest level of education completed?	Texto	Mayor nivel educativo completado.
What is your gender?	Texto	Genero
What is your race? (Choose all that apply.)	Texto	Raza

2. Modelado de datos – procesamiento y transformación de los datos

El modelado de los datos realizado en Python tuvo en cuenta varias etapas:

En la primera etapa se realizó la extracción de los datos de la fuente original (https://docs.google.com/spreadsheets/d/1IPS5dBSGtwYVbjsfbaMCYIWnOuRmJcbequohNxCyGVw/edit?resourcekey#gid=1625408792) y el cargue de los datos para su procesamiento. En esta etapa también se tiene la primera aproximación a los datos cargados.

En la segunda etapa se realizó la identificación de las variables y sus tipos (object, float, int) y se estandarizaron los nombres para ser mas eficientes en el procesamiento. A continuación, se presenta la relación entre los nombres de la base original y los nuevos nombres asignados:

Nombre de la variable - original	Nombre de la variable - nuevo
Timestamp	Fecha
How old are you?	Edad
What industry do you work in?	Industria
Job title	Cargo
If your job title needs additional context, please clarify here	Contexto cargo
What is your annual salary? (You'll indicate the currency in a later question. If you are part-time or hourly, please enter an annualized equivalent -- what you would earn if you worked the job 40 hours a week, 52 weeks a year.)	Salario anual
How much additional monetary compensation do you get, if any (for example, bonuses or overtime in an average year)? Please only include monetary compensation here, not the value of benefits.	Otras compensaciones
Please indicate the currency	Moneda
If "Other," please indicate the currency here	Otra moneda
If your income needs additional context, please provide it here	Contexto salario
What country do you work in?	Pais
If you're in the U.S., what state do you work in?	Estado (USA)
What city do you work in?	Ciudad
How many years of professional work experience do you have overall?	experiencia
How many years of professional work experience do you have in your field?	experiencia especifica
What is your highest level of education completed?	Educacion
What is your gender?	Genero
What is your race? (Choose all that apply.)	Raza

En la tercera etapa se realizó la transformación de la variable 'pais', 'ciudad' y 'moneda' ('Otra moneda') para homogeneizar las categorías. Además, se crean los campos de 'salario_COP' (salario anual en pesos colombianos) y 'Compensaciones_COP' (compensaciones en pesos colombianos) utilizando la tasa de cambio de cada moneda a pesos colombianos. Por último, se crea el 'Ingreso_total' como la suma del salario y las compensaciones en pesos colombianos.

3. Esquema de trabajo – procesamiento y transformación de los datos

En caso de que se requiera actualizar los datos por una versión más reciente o solo replicar el ejercicio, se deben seguir los siguientes pasos:

- Descargar la base de datos actualizada: descargar el archivo en formato .csv y nombrarlo '*Ask A Manager Salary Survey 2021 (Responses) - Form Responses 1.csv*' (esto ultimo solo si se quiere aprovechar el código desarrollado en Python).
- Cargar la base de datos en Python (este desarrollo se hizo en notebooks de jupyter pero pueden utilizar la interfaz que mejor les parezca). Se utilizó la función 'read_csv' de la librería pandas.
- Renombrar variables: a través de un diccionario que contiene los nombres de las variables originales de la base de datos y los nuevos nombres que se van a asignar a las columnas/variables, se implementó la función 'rename' para codificar las variables con los nuevos nombres. La tabla presentada en el punto 2 de este documento contiene las llaves y los valores para implementar el diccionario en Python.
- Limpiar los campos 'Pais' y 'Ciudad': para realizar la limpieza de los datos se utilizó la función 'replace' y expresiones regulares 'regex'. Primero se eliminan los espacios en blanco al principio y al final de la cadena de caracteres (valor que toma el campo, ejemplo: ' Colombia '), seguido, se estandarizan todos los nombres a minúsculas para facilitar su homogeneización. Tercero, se eliminan los puntos y se reemplazan los nombres de los países y ciudades utilizando expresiones regulares, ejemplo: Estados Unidos se encontraba escrito de muchas maneras como 'USA', 'US', 'u s', 'United states', 'United States America', y se emplean expresiones regulares para expresar patrones que permitan reemplazar esos nombres por 'Estados Unidos'. En este proceso se imprime por ventana la lista de los países ordenada para validar de una manera mas sencilla la similitud entre 2 o más países o ciudades.
- Se creó la variable 'Moneda_nueva' en la que se incluyen las monedas que se encuentran en el listado de la encuesta y las nuevas monedas para los usuarios que seleccionaron la opción 'otra'. Este proceso también requirió de la estandarización de las monedas, que resultaron en 38 tipos de monedas, a través de la función 'replace' y 'regex' con el mismo procedimiento explicado en el apartado anterior.
- Finalmente, se realiza la consulta manual de los tipos de cambio para cada uno de los 38 tipos de monedas en https://www.xe.com/currencyconverter/ para crear un diccionario (las llaves son los tipos de monedas y los valores el tipo de cambio). Se crea la variable 'Tasa_cambio' en la base de datos y se multiplica el salario anual por el tipo de cambio para crear el campo/variable 'Salario_COP' y 'Compensaciones_COP'.

Este ejercicio termina con la base de datos exportada en formato '.csv' y su carga a Power BI (o la herramienta de visualización que estimen conveniente).