## Informe Modelado de Variables: Ask A Manager Salary Survey 2021 (Responses)

1. Variables en base de datos original.

| Nombre | Nombre del Campo Original | Tipo de Variable | Descripción |
|---|---|---|---|
| Edad | How old are you? | Texto | Seleccionar el rango de edades:<br>under 18<br>18-24<br>25-34<br>35-44<br>45-54<br>55-64<br>65 or over |
| Industria | What industry do you work in? | Texto | Selección múltiple de diferentes industrias tales como, campo opcional:<br>Accounting, Banking & Finance<br>Agriculture or Forestry<br>Art & Design<br>Business or Consulting<br>Computing or Tech<br>Education (Primary/Secondary)<br>Other:<br>etc. |
| Cargo | Job title | Texto | Especifica el cargo del encuestado |
| Contexto Cargo | If your job title needs additional context, please clarify here: | Texto | Especifica si el cargo del encuestado requiere algún contexto adicional, este campo es opcional |
| Salario Anual | 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.) | Numérico | Especificar el salario anual |
| Otras Compensaciones | 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. | Numérico | Especificar compensaciones adicionales, campo opcional |
| Moneda (Currency) | Please indicate the currency | Texto | Especificar el tipo de moneda bajo el cual puso el salario anual dentro de una lista de opciones, por ejemplo, USD, EUR, etc. |
| Moneda-Otra | If "Other," please indicate the currency here: | Texto | Si el encuestado en la pregunta de indicar el tipo de moneda selecciono Other, en esta pregunta deberá especificar el tipo de moneda, campo opcional. |

| | | | |
|---|---|---|---|
| Contexto Ingresos | If your income needs additional context, please provide it here: | Texto | Especificar si el ingreso necesita contexto adicional, campo opcional |
| Pais- Trabajo | What country do you work in? | Original: Texto<br>Modificación: Geográfico - País | País en el que trabaja |
| USA - Estado | If you're in the U.S., what state do you work in? | Texto | Si trabaja en Estados Unidos de la lista de opciones especifique el Estado, campo opcional |
| Ciudad- Trabajo | What city do you work in? | Original: Texto<br>Modificación: Geográfico - País | Ciudad en la que trabaja |
| Años de Experiencia | How many years of professional work experience do you have overall? | Texto | Seleccionar los años de experiencia profesional de la siguiente lista de opciones:<br>1 year or less<br>2 - 4 years<br>5-7 years<br>8 - 10 years<br>11 - 20 years<br>21 - 30 years<br>31 - 40 years<br>41 years or more |
| Años de Experiencia en el sector | How many years of professional work experience do you have in your field? | Texto | Seleccionar los años de experiencia profesional en el sector/ campo de la siguiente lista de opciones:<br>1 year or less<br>2 - 4 years<br>5-7 years<br>8 - 10 years<br>11 - 20 years<br>21 - 30 years<br>31 - 40 years<br>41 years or more |
| Educación | What is your highest level of education completed? | Texto | Seleccionar el nivel más alto de educación alcanzado por el encuestado, campo opcional:<br>Master's degree<br>College degree<br>PhD<br>Some college<br>High School<br>Professional degree (MD, JD, etc.) |

| | | | |
|---|---|---|---|
| Genero | What is your gender? | Texto | Especificar el género al que pertenece.<br>Campo opcional:<br>Man<br>Woman<br>Non-binary<br>Other or prefer not to answer |
| Raza | What is your race? (Choose all that apply.) | Texto | Seleccionar la raza a la que pertenece en donde puede seleccionar todas las opciones que aplican, campo opcional:<br>Asian or Asian American<br>Black or African American<br>Hispanic, Latino, or Spanish origin<br>Middle Eastern or Northern African<br>Native American or Alaska Native<br>White<br>Another option not listed here or prefer not to answer |

2. Variables agregadas para el modelado

| | | Descripción | Tipo de variable |
|---|---|---|---|
| 18 | Pais | En este campo se realiza la estandarización de los países | Geográfico - Pais |
| 19 | Ciudades | En este campo se realiza la estandarización de las ciudades | Geográfico - Ciudad |
| 20 | Salario Modificado | En este campo los valores que en Salario Anual son menores que 100 se asume que se deben multiplicar por 10000 | Numérico |
| 21 | Salario Anual COP | Conversión del campo Salario Modificado a pesos colombianos teniendo en cuenta el tipo de moneda que el usuario indico en el campo Moneda(currency) | Moneda COP |
| 22 | Compensaciones COP | Conversión del campo Otras Compensaciones a pesos colombianos teniendo en cuenta el tipo de moneda que el usuario indico en el campo Moneda(currency) | Moneda COP |
| 23 | Salario COP | Suma del Salario Anual COP + Compensaciones COP | Moneda COP |

3. Paso a Paso actualización de los datos

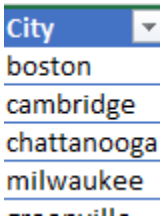
**Estandarización países**: En Looker usando la opción de Añadir un Campo, a continuación, encontrara el código usado para la estandarización de los Países en donde inicialmente se eliminan los espacios y se convierte en minúscula los textos que ingresaron los encuestados, posteriormente se agrupan las diferentes maneras como están escritos los países y se define el nombre bajo el cual quedara finalmente, es necesario tener en cuenta que aquellos registros que no son países se convierten a NULL.

```
CASE
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "united  states","u.s.", "united states", "us", "us ", "america", "the united states", "the us", "u. s", "u. s.", "u.a.", "u.s", "u.s.", "u.s>", "u.sa", "ua", "u.s.a.", "usa", "u.s.a", "uxz", "united y", "united statss", "uniteed states", "usa (company is based in a us territory, i work remote)", "usa-- virgin islands", "usa, but for foreign gov't", "uniited states", "unite states", "united states", "united states of america", "united states- puerto rico", "unites states", "usa tomorrow", "unitef stated", "united states of american", "united stares", "united state", "united state of america", "united stated", "united stateds", "united states is america", "unitedstates", "united stattes", "united statesp", "unted states", "for the united states government, but posted overseas", "us govt employee overseas, country withheld", "us of a", "usa", "usa tomorrow", "usa-- virgin islands", "usa, but for foreign gov't", "usaa", "usab", "usat", "usd", "uss", "uxz", "california", "united statea", "us", "united status", "san francisco", "united sates", "united  sates") THEN "USA"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "united sates", "united sates of america", "united stares", "united state", "united state of america", "united statea", "united stated", "united stateds", "united statees", "united states", "united states (i work from home and my clients are all over the us/canada/pr", "united states is america", "united states of america", "united states of american", "united states of americas", "united statesp", "united statew", "united statss", "united stattes", "united statues", "united status", "united statws", "united sttes", "united y", "unitedstates", "uniteed states", "unitef stated", "uniter statez", " us", "us ", "hartford", "i.s.", "i work for a uae-based organization, though i am personally in the us.", "isa", "unitied states", "untied states", "united states", "uniyes states", "uniyed states" ) THEN "USA"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "united kingdom","united kindom", "uk", "england", "england, united kingdom","england, gb", "england, uk", "england, uk.", "england/uk", "englang", "england, uk", "great britain", "london", "scotland", "scotland, uk", "u.k", "u.k.", "u.k. (northern england)", "uk (england)", "uk (northern ireland)", "uk for u.s. company", "uk, but for globally fully remote company", "uk, remote", "united kingdom (england)", "united kingdom.", "united kingdomk", "britain", "wales", "northern ireland", "northern ireland, united kingdom", "wales", "wales (uk)", "wales (united kingdom)", "wales, uk", "unites kingdom" ) THEN "UK"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "australi", "australia", "australian" ) THEN "Australia"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "can", "canad", "canada", "canadá", "canada and usa", "canada, ottawa, ontario", "canadw", "canda", "csnada", "i am located in canada but i work for a company in the us" ) THEN "Canada"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "danmark", "dbfemf", "denmark" ) THEN "Denmark"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "new zealand", "new zealand aotearoa", "nz", "from new zealand but on projects across apac" ) THEN "New Zealand"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "nederland", "netherlands", "the netherlands" ) THEN "Netherlands"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "argentina but my org is in thailand", "argentina", "i work for an us based company but i'm from argentina." ) THEN "Argentina"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ("company in germany. i work from pakistan.") THEN "Pakistan"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "brasil", "brazil" ) THEN "Brasil"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ("from romania, but for an us based company") THEN "Romania"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "hong kong", "hong konh" ) THEN "Hong Kong"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "japan", "japan, us gov position" ) THEN "Japan"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "italy", "italy (south)" ) THEN "Italy"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "luxembourg", "luxemburg" ) THEN "Luxembourg"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "austria, but i work remotely for a dutch/british company", "austria" ) THEN "Austria"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "czech republic", "czechia" ) THEN "Czech Republic"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "mexico", "méxico" ) THEN "Mexico"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ("remote (philippines)") THEN "Phillippines"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ("germany") THEN "Germany"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ("ireland") THEN "Ireland"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ("mainland china") THEN "China"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ("croatia") THEN "Croatia"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ("switzerland") THEN "Switzerland"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ("spain") THEN "Spain"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "ibdia", "india" ) THEN "India"
  WHEN TRIM(LOWER(Pais-Trabajo)) IN ( "y", "we don't get raises, we get quarterly bonuses, but they periodically asses income in the area you work, so i got a raise because a 3rd party assessment showed i was paid too little for the area we were located", "worldwide (based in us but short term trips aroudn the world)", "contracts", "$2,175.84/year is deducted for benefits", "bonus based on meeting yearly goals set w/ my supervisor", "i earn commission on sales. if i meet quota, i'm guaranteed another 16k
```

```
min. last year i earned an additional 27k. it's not uncommon for people in my space to earn 100k+ after commission.", "i was brought in on this salary to help with the ehr and very quickly was promoted to current position but compensation was not altered.", "currently finance", "ff", "i earn commission on sales. if i meet quota, i'm guaranteed another 16k min. last year i earned an additional 27k. it's not uncommon for people in my space to earn 100k+ after commission.", "i was brought in on this salary to help with the ehr and very quickly was promoted to current position but compensation was not altered.", "loutreland", "remote", "international", "global", "n/a (remote from wherever i want)", "na", "policy", "ss" ) THEN NULL
  ELSE TRIM(Pais-Trabajo)
END
```

**Estandarización ciudades**: Debido a la gran cantidad de registros de ciudades en la base a continuación se encuentra el paso a paso realizado para la estandarización de los textos de las ciudades:

1. Copiar las ciudades de la base en otra hoja de Excel
2. A esta lista aplicar la formula para quitar espacios al inicio y fin de los textos y convertir el texto en minúsculas (ESPACIOS(MINUSC())
3. Copiar estos valores como texto en un nuevo archivo de Excel y eliminar los duplicados el nombre de la columna se definió como City y el archivo se guardó como ciudades




4. El siguiente código en Python descarga una lista de ciudades válidas desde una fuente externa, luego procesa un archivo Excel (que en el paso 4 guardamos como ciudades) que contiene los diferentes textos de las ciudades. Establece un formato estándar para los nombres de las ciudades, corrige errores tipográficos usando coincidencia difusa, y filtra las ciudades no válidas. Finalmente, guarda el archivo con los nombres de ciudades estandarizados.

```
import pandas as pd
import zipfile
import requests
from rapidfuzz import process

def get_free_city_list():
  """
  Obtiene una lista de ciudades válidas desde una fuente gratuita como GeoNames o SimpleMaps.
  """
  url = "https://simplemaps.com/static/data/world-cities/basic/simplemaps_worldcities_basicv1.75.zip"

  # Descargar y extraer el archivo ZIP
  zip_file_path = "simplemaps_worldcities_basicv1.75.zip"
  with requests.get(url) as r:
    with open(zip_file_path, 'wb') as f:
      f.write(r.content)

  # Extraer el archivo Excel del ZIP
  with zipfile.ZipFile(zip_file_path, 'r') as zip_ref:
    zip_ref.extract('worldcities.xlsx')  # Especificar el archivo Excel

  # Leer el archivo Excel con pandas
```

```
    df_cities = pd.read_excel('worldcities.xlsx', usecols=[0])  # Suponiendo que la primera columna tiene los nombres de ciudades
    return df_cities["city"].dropna().str.title().tolist()

def is_valid_city(city_name, valid_city_list):
    """
    Verifica si el nombre es una ciudad válida comparándolo con una lista de ciudades conocidas.
    """
    result = process.extractOne(city_name, valid_city_list, score_cutoff=85)

    # Verificar si se encontró un match
    if result:
        match, score, _ = result  # extraemos solo los dos primeros valores (match y score)
        return match is not None
    return False

def standardize_city_names(df, column_name, valid_city_list):
    """
    Estandariza los nombres de ciudades en una columna específica de un DataFrame.
    - Convierte a formato título (capitalizando cada palabra).
    - Elimina espacios innecesarios.
    - Corrige nombres similares usando coincidencia difusa.
    - Filtra registros que no sean ciudades válidas.
    - Conserva la columna con el nombre original.
    """
    df["Original_" + column_name] = df[column_name]  # Guardar la columna original
    df[column_name] = df[column_name].str.strip().str.title()

    # Filtrar las ciudades válidas
    df = df[df[column_name].apply(lambda city: is_valid_city(city, valid_city_list))]

    unique_cities = df[column_name].unique()
    standardized_names = {}

    for city in unique_cities:
        result = process.extractOne(city, standardized_names.keys(), score_cutoff=85)
        if result:
            match, _, _ = result  # Extraer solo el match
            standardized_names[city] = standardized_names[match]
        else:
            standardized_names[city] = city

    df[column_name] = df[column_name].map(standardized_names)
    return df

# Cargar el archivo Excel
df = pd.read_excel("ciudades.xlsx")

# Obtener lista de ciudades válidas de una fuente gratuita
valid_city_list = get_free_city_list()

# Aplicar la estandarización y filtrado
df = standardize_city_names(df, "City", valid_city_list)
```

```
# Guardar el archivo con los nombres estandarizados
df.to_excel("ciudades_estandarizadas.xlsx", index=False)

print("Archivo procesado y guardado como 'ciudades_estandarizadas.xlsx'")
```

5. Una vez obtenido este nuevo archivo con la estandarizacion de las cuidades se crea una columna adicional en la base de datos y con un buscarv poder traer la ciudad estandarizada, en caso de que no encuentre la ciudad este campo se pondrá como null

**Salario Modificado:** Este campo añadido usa la siguiente formula en Looker con el fin de poder multiplicar todos aquellos valores menores a 100 que se asume son valores que pueden estar refiriéndose a coloquialismos como 80K

```
CASE
  WHEN Salario Anual < 100 THEN Salario Anual * 10000
  ELSE Salario Anual
END
```

**Salario Anual COP**: Este campo añadido a partir de la siguiente formula convierte a pesos colombianos los salarios que ingresaron los encuestados usando la TRM extraída manualmente de https://www.xe.com/currencyconverter/

```
CASE
  WHEN Moneda (Currency) = "USD" THEN (Salario Modificado) * 4167.45
  WHEN Moneda (Currency) = "CAD" THEN (Salario Modificado) * 2890.89
  WHEN Moneda (Currency) = "GBP" THEN (Salario Modificado) * 5188.89
  WHEN Moneda (Currency) = "EUR" THEN (Salario Modificado) * 4344.23
  WHEN Moneda (Currency) = "AUD/NZD" THEN (Salario Modificado) * 2598.91
  WHEN Moneda (Currency) = "CHF" THEN (Salario Modificado) * 4598.32
  WHEN Moneda (Currency) = "SEK" THEN (Salario Modificado) * 379.10
  WHEN Moneda (Currency) = "JPY" THEN (Salario Modificado) * 26.99
  WHEN Moneda (Currency) = "ZAR" THEN (Salario Modificado) * 224.74
  WHEN Moneda (Currency) = "HKD" THEN (Salario Modificado) * 534.84
  WHEN Moneda (Currency) = "Other" AND Moneda-Otra = "USD" THEN (Salario Modificado) * 4167.45
  ELSE (Salario Modificado)
END
```

**Compensaciones COP:** Similar a la formula de Salario Anual COP, en este campo se multiplica el valor que el usuario ingreso en el campo de Otras compensaciones por la TRM extraída manualmente de https://www.xe.com/currencyconverter/

```
CASE
  WHEN Moneda (Currency) = "USD" THEN  Otras compensaciones * 4167.45
  WHEN Moneda (Currency) = "CAD" THEN  Otras compensaciones * 2890.89
  WHEN Moneda (Currency) = "GBP" THEN Otras compensaciones * 5188.89
  WHEN Moneda (Currency) = "EUR" THEN Otras compensaciones * 4344.23
  WHEN Moneda (Currency) = "AUD/NZD" THEN Otras compensaciones * 2598.91
  WHEN Moneda (Currency) = "CHF" THEN Otras compensaciones * 4598.32
  WHEN Moneda (Currency) = "SEK" THEN Otras compensaciones * 379.10
  WHEN Moneda (Currency) = "JPY" THEN Otras compensaciones * 26.99
  WHEN Moneda (Currency) = "ZAR" THEN Otras compensaciones * 224.74
```

```
  WHEN Moneda (Currency) = "HKD" THEN Otras compensaciones * 534.84
  WHEN Moneda (Currency) = "Other" AND Moneda-Otra = "USD" THEN Otras compensaciones * 4167.45
  ELSE Otras compensaciones
END
```

Salario COP: Este campo añadido suma del Salario Anual COP + Compensaciones COP

Salario Anual COP+Compensaciones COP

4.