**Universidad Autónoma Nacional de México**

Facultad de Estudios Superiores Acatlán

# Diplomado en Ciencia de Datos

4ª Generación

**Efraín Ismael Flores Hernández**

## Examen 1 Módulo I

septiembre, 2020

## Índice

## Introducción

*“Cuando uno busca algo con la certeza de que va a encontrarlo, es muy probable que lo encuentre. Aunque no sea lo que realmente buscaba.”*

*John Katzenbach*

El conjunto de datos de este examen contiene información mensual del número de asegurados por el Instituto Mexicano de Seguridad Social (IMSS) junto con el salario base total, el desglose ocurre para las siguientes dimensiones:

- Entidad y municipio
- Delegación y subdelegación de adscripción
- Sector socioeconómico (jerárquicamente a 3 niveles)
- Tamaño del registro patronal
- Sexo y rango de edad
- Rango salarial con base en salarios mínimos y con base en la Unidad de Medida de Actualización (UMA)

Para este examen, las instrucciones constan en conocer la calidad de los datos (más de 4 millones de registros), los desafíos que presentan y una propuesta de cómo utilizar inteligencia de negocio y analítica avanzada en ellos para resolver un problema.

## Calidad de datos

Primero, se describirán las variables junto con su tipo:

**Tabla 1. Descripción de variables**

| # | Columna | Descripción | Tipo | Natualeza del dato |
|---|---|---|---|---|
| 1 | cve_delegacion | Identificador de la delegación de adscripción al IMSS | Numérico | Dimensión |
| 2 | cve_subdelegacion | Identificador de la subdelegación de adscripción al IMSS | Numérico | Dimensión |
| 3 | cve_entidad | Identificador de la entidad federativa (catálogo IMSS) | Numérico | Dimensión |
| 4 | cve_municipio | Identificador del municipio (catálogo IMSS) | Alfanumérico | Dimensión |
| 5 | sector_economico_1 | Identificador de sector económico a 1 posición (división, catálogo IMSS) | Numérico | Dimensión |
| 6 | sector_economico_2 | Identificador de sector económico a 2 posiciones (grupo, catálogo IMSS) | Numérico | Dimensión |
| 7 | sector_economico_4 | Identificador de sector económico a 4 posiciones (fracción, catálogo IMSS) | Numérico | Dimensión |
| 8 | tamaño_patron | Identificador del rango de tamaño del registro patronal | Alfanumérico | Dimensión |
| 9 | sexo | 1: Hombre 2: Mujer | Numérico | Dimensión |
| 10 | rango_edad | Identificador de rango de edad | Alfanumérico | Dimensión |
| 11 | rango_salarial | Identificador de rango salarial | Alfanumérico | Dimensión |
| 12 | rango_uma | identificador del rango de la unidad de medida de actualización (UMA). A partir de febrero de 2017 | Alfanumérico | Dimensión |
| 13 | asegurados | Número de asegurados | Numérico | Métrica |
| 14 | no_trabajadores | Asegurados sin un empleo asociado | Numérico | Métrica |
| 15 | ta | Puestos de trabajo afiliados al IMSS (empleos asegurados o asegurados asociados a un empleo) | Numérico | Métrica |
| 16 | teu | Puestos de trabajo eventuales urbanos | Numérico | Métrica |
| 17 | tec | Puestos de trabajo eventuales del campo | Numérico | Métrica |
| 18 | tpu | Puestos de trabajo permanentes urbanos | Numérico | Métrica |
| 19 | tpc | Puestos de trabajo permanentes del campo | Numérico | Métrica |
| 20 | ta_sal | Puestos de trabajo afiliados con un salario asociado | Numérico | Métrica |
| 21 | teu_sal | Puestos de trabajo eventuales urbanos con un salario asociado | Numérico | Métrica |
| 22 | tec_sal | Puestos de trabajo eventuales del campo con un salario asociado | Numérico | Métrica |
| 23 | tpu_sal | Puestos de trabajo permanentes urbanos con un salario asociado | Numérico | Métrica |
| 24 | tpc_sal | Puestos de trabajo permanentes del campo con un salario asociado | Numérico | Métrica |
| 25 | masa_sal_ta | Masa salarial asociada a puestos de trabajo afiliados | Numérico | Métrica |
| 26 | masa_sal_teu | Masa salarial asociada a puestos de trabajo eventuales urbanos | Numérico | Métrica |
| 27 | masa_sal_tec | Masa salarial asociada a puestos de trabajo eventuales del campo | Numérico | Métrica |
| 28 | masa_sal_tpu | Masa salarial asociada a puestos de trabajo permanentes urbanos | Numérico | Métrica |
| 29 | masa_sal_tpc | Masa salarial asociada a puestos de trabajo permanentes del campo | Numérico | Métrica |

*Fuente: Diccionario de datos IMSS*

Se analizará la calidad de nuestro conjunto de datos con la siguiente estructura:

- Valores ausentes
- Valores Atípicos
- Catálogos

## Ausentes

La variable con más registros vacíos es la #4, la clave de municipio. Según el documento con nombre: “Preguntas frecuentes de asegurados”, que el registro esté ausente significa que la entidad es la Ciudad de México, no cuenta con municipios. Entonces, basta con reemplazar los municipios vacíos por “MEX” para distinguirlos.

Respecto a las variables del sector económico, los valores ausentes significan que no hay empleo asociado, pueden ser asegurados familiares de trabajadores en el IMSS, CFE, etc. No existe riesgo al omitirlos dado que sólo es el 0.36% de los registros totales.

Ahora para la variable #11, cuando el rango salarial está vacío, el registro puede agruparse con la categoría mínima: “W1: hasta 1 salario mínimo”. Aún así es la categoría con menos registros. El mismo tratamiento se aplica para la variable: “rango_uma”.

Para la variable restante con registros vacíos es la que define el tamaño del registro patronal, los vacíos también implican que los asegurados no tienen un empleo asociado, de la misma manera que el sector económico, también se opta por omitirlos. Como dato curioso, de los registros vacíos en esta variable, el 65.5% de ellos corresponden al sector económico de “servicios domésticos” y el resto “servicios profesionales y técnicos”.

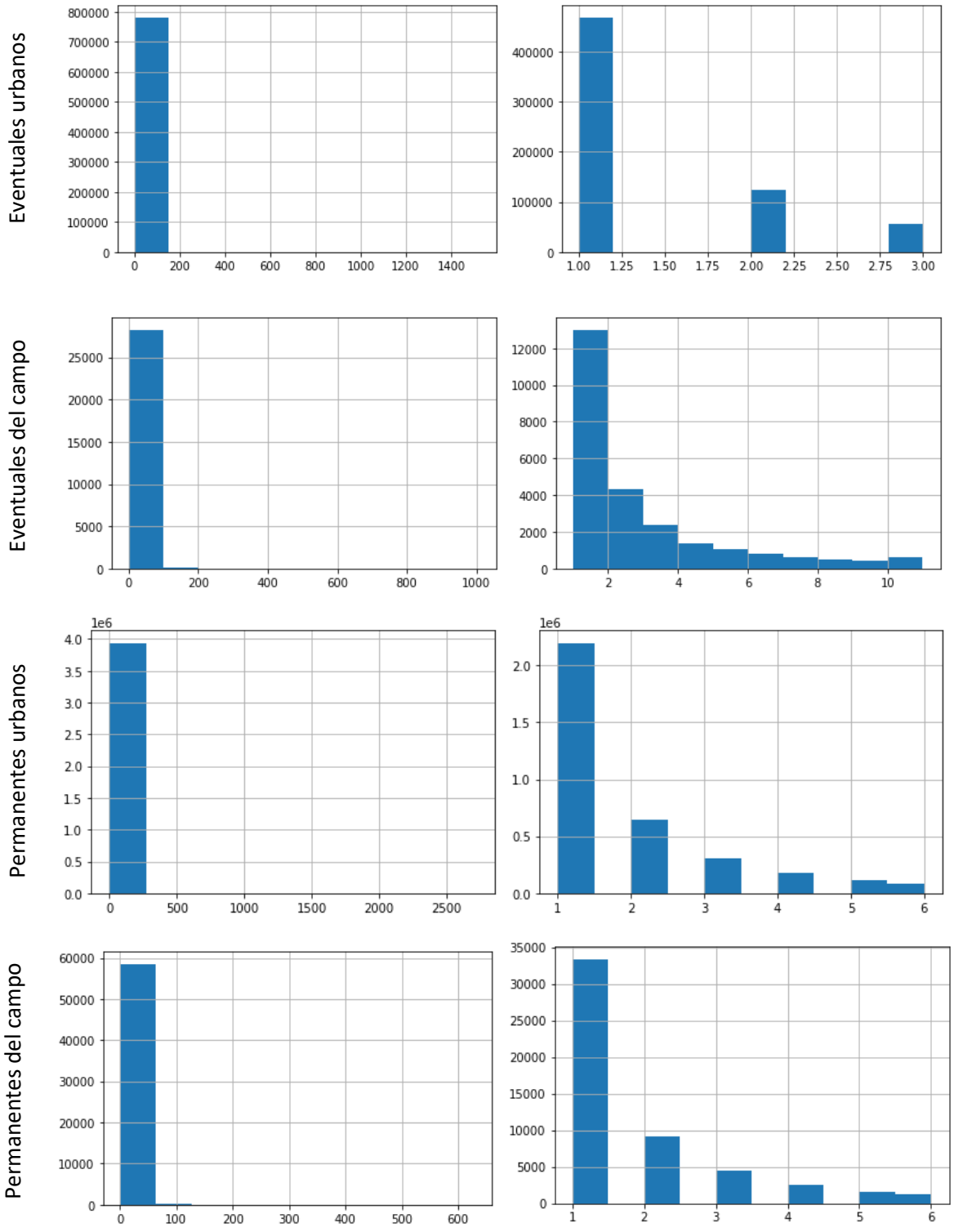
En cuanto a la limpieza de valores ausentes, la tabla está lista para proceder con la limpieza de valores atípicos. Continuamos con el 99.4% de los registros originales.

## Atípicos

Al parecer, la cantidad de asegurados es igual a la variable #15, que despliega el número de puestos de trabajo afiliados al IMSS, es una variable redundante, se elimina. De manera similar, la variable llamada “no_trabajadores” sólo presenta el valor 0 para todos los registros, es irrelevante y, por lo tanto, también se omiten.

Dado que la variable “ta” es la suma de los puestos eventuales y permanentes tanto urbanos como los del campo, se procede a evaluar los valores atípicos en cada uno de los tipos y temporalidad de puestos.

**Gráfica 1. Puestos por tipo y temporalidad, antes y después de tratar outliers**



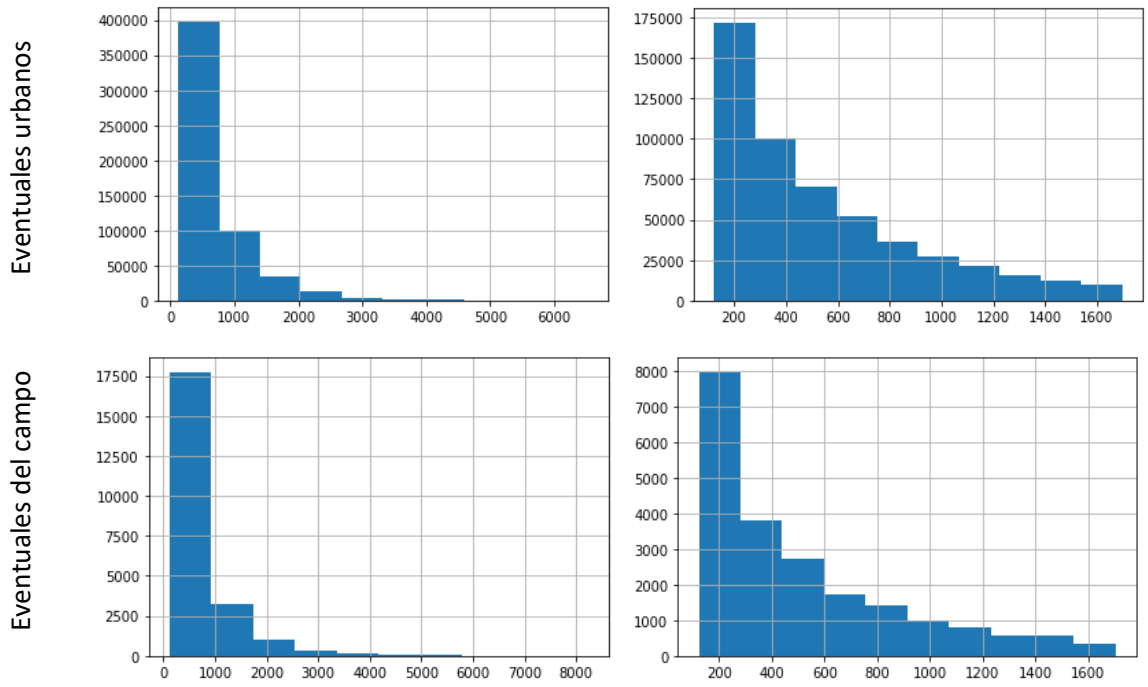

*Fuente: Elaboración propia con el conjunto de datos*

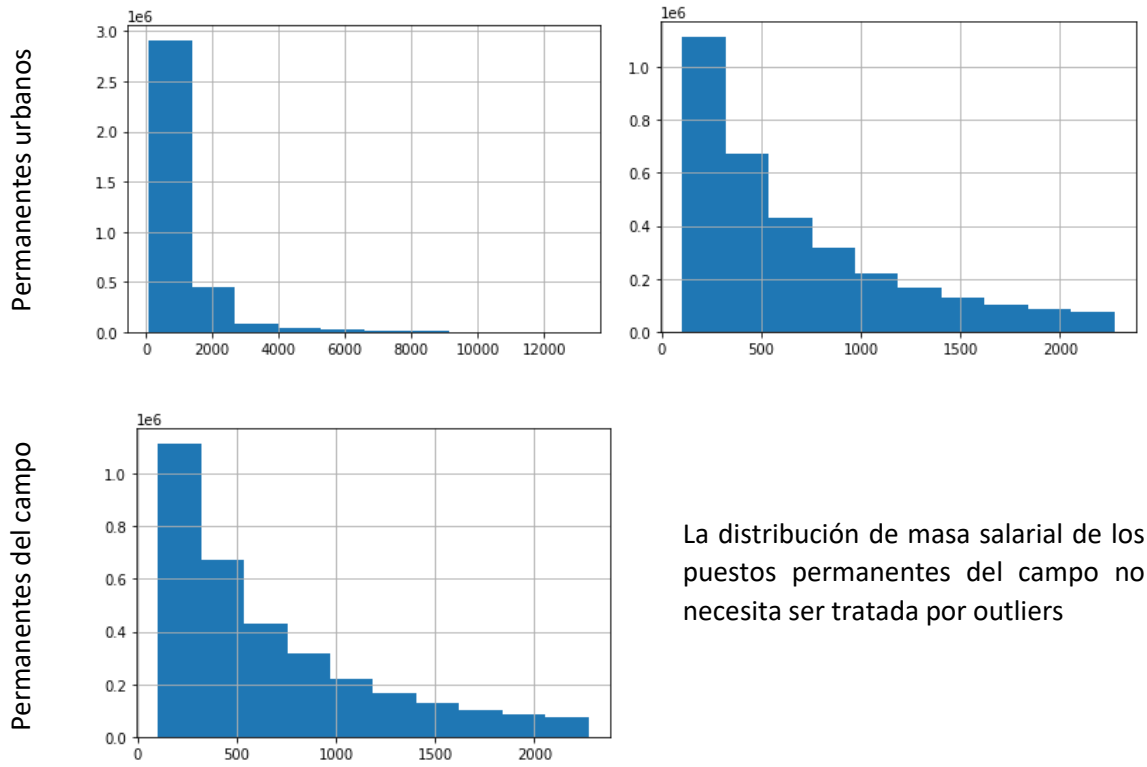
Dado que cada uno de los componentes de la variable “ta” no tienen outliers, la distribución de la suma es más cercana a la realidad. Aún contemplamos el 87.2% de los registros originales.

¿Existen puestos afiliados sin salario asociado? Al parecer sí, pero son menos del 0.2% de los registros, de éstos el 73% se refiere a un solo puesto respecto al tamaño de registro patronal. Además, casi la mitad de estos registros pertenecen al sector comercio. Se opta por omitir los registros con puestos sin salario asociado, continuamos con el 87% de registros aún disponibles. Entonces, se vuelve redundante conservar las columnas que describen los puestos con salario asociado, se eliminan.

De la misma manera que se trataron los valores atípicos en las variables que definen la cantidad de puestos por tipo y temporalidad, se trabajan los valores atípicos de las variables que despliegan la masa salarial.

**Gráfica 2. Masa salarial por tipo y temporalidad, antes y después de tratar outliers**






La distribución de masa salarial de los puestos permanentes del campo no necesita ser tratada por outliers

*Fuente: Elaboración propia con el conjunto de datos*

## Catálogos

En el archivo de diccionario de datos, se describen detalladamente las claves de cada variable categóricas. Es importante revisar que estos catálogos estén correctamente estructurados.

El identificador (ID) de subdelegación depende del ID de la delegación, por ejemplo: tanto la delegación 1 y 2 tienen una subdelegación identificada como 1. Entonces, si el catálogo define a Mexicali con el ID 1, y esta subdelegación pertenece a la delegación Baja California con el ID 2, entonces el ID real de la subdelegación será construido como 02_01. Con esta interpretación, se define el ID real de cada subdelegación.

**Tabla2. Muestra de catálogo para subdelegación**

| ID_subdeleg | Subdelegación |
|---|---|
| 05_03 | Saltillo |
| 07_01 | Tuxtla Gutiérrez |
| 27_57 | Nacozari de García |
| 39_57 | Centro |

*Fuente: Elaboración propia con el diccionario de datos*

De esta forma, se pueden separar los catálogos de delegación y subdelegación. Ambos sin duplicados y con ID única.

Con el catálogo de entidad y municipio, un tratamiento similar al anterior puede ser aplicado. En este caso, existen claves de municipio pertenecientes a la entidad CDMX, pero sin nombre de municipio, no es necesario contar con estas claves así que se opta por omitirlas.

La clave de municipio consta de 3 caracteres alfanuméricos, sin embargo, al comprobar que el ID es único, existe la clave “Y44” duplicada, referente al municipio de Puebla, Puebla y también al municipio “General Plutarco Elías Calles”, Sonora. Entonces, se rea el ID real de municipio uniendo la clave de entidad a dos dígitos concatenado la clave alfanumérica de municipio. Por ejemplo:

**Tabla3. Muestra de catálogo para municipio**

| ID_municipio | Municipio |
|---|---|
| 21_Y44 | Puebla |
| 26_Y44 | General Plutarco Elías Calles |
| 20_V56 | San Carlos Yautepec |

*Fuente: Elaboración propia con el diccionario de datos*

Dado que en nuestro conjunto de datos reemplazamos los registros vacíos en la clave de municipio por "MEX" y dado que en nuestro catálogo eliminamos los registros vacíos, tendríamos que crear un registro con la clave de entidad 9 (perteneciente a CDMX) y la clave de municipio "MEX" para traer el nombre "CDMX".

Para los catálogos restantes, la idea es estructurarlos en dos columnas: ID y descripción, para trabajarlos como objetos de tipo diccionario y así reemplazarlo en el conjunto de datos. Afortunadamente no ocuparon reestructuración como los mencionados anteriormente, bastó con cambiar a mayúsculas el catálogo de tamaño de registro patronal para que empatara con la variable en la tabla principal.

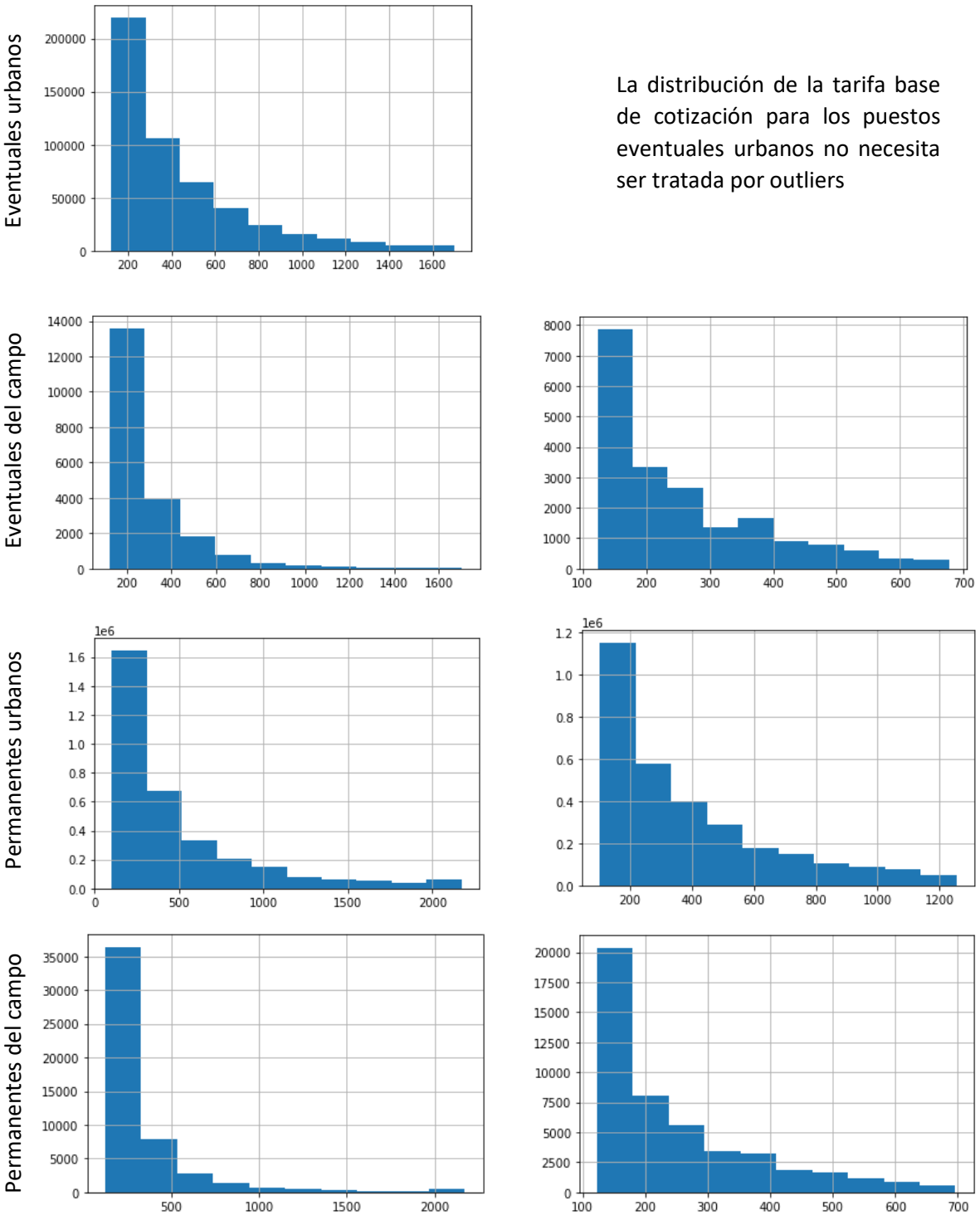
**Tabla4. Ejemplo de registro con catálogo**

| Columna | Ejemplo |
|---|---|
| cve_delegacion | Estado de México Poniente |
| cve_subdelegacion | Naucalpan |
| cve_entidad | Estado México |
| cve_municipio | Naucalpan de Juárez |
| sector_economico_1 | Industrias de transformación |
| sector_economico_2 | Construcción, reconstrucción y ensamble de equ... |
| sector_economico_4 | Fabricación y/o ensamble de partes y accesorio... |
| tamaño_patron | con 51 y hasta 250 puestos de trabajo |
| sexo | Hombre |
| rango_edad | Mayor o igual a 30 y menor a 35 años de edad |
| rango_salarial | mayor a 10 y hasta 11 veces el salario mínimo |
| rango_uma | mayor a 10 y hasta 11 veces la UMA |
| ta | 1 |
| teu | 0 |
| tec | 0 |
| tpu | 1 |
| tpc | 0 |
| masa_sal_ta | 588.77 |
| masa_sal_teu | 0 |
| masa_sal_tec | 0 |
| masa_sal_tpu | 588.77 |
| masa_sal_tpc | 0 |

*Fuente: Elaboración propia con el diccionario de datos*

Se cuenta con la masa salarial y el número de puestos afiliados, con la razón entre estas dos variables se calcula la tarifa base de cotización, que significa cuánto gana un trabajador al día, con base en ello se calculan sus prestaciones. Esta nueva variable también es candidata para tratar outliers.

**Gráfica 3. Tarifa base de cotización por tipo y temporalidad, antes y después de tratar outliers**




La distribución de la tarifa base de cotización para los puestos eventuales urbanos no necesita ser tratada por outliers

*Fuente: Elaboración propia con el conjunto de datos*

## Desafíos

Además de los desafíos de limpieza ya aplicados, la estructura actual del conjunto de datos puede no ser práctica para análisis entre tipos y temporalidad de puestos (eventuales y permanentes tanto urbanos como del campo). Así que, a través de la función “melt” de la paquetería pandas para Python la estructura cambia:

**Tabla5. Ejemplo de registro con nueva estructura**

| Columna | Ejemplo |
|---|---|
| cve_delegacion | Guanajuato |
| cve_subdelegacion | León |
| cve_entidad | Guanajuato |
| cve_municipio | León |
| sector_economico_1 | Servicios para empresas, personas y el hogar |
| sector_economico_2 | Servicios de alojamiento temporal ... |
| sector_economico_4 | Servicios de alojamiento temporal |
| tamaño_patron | con 51 y hasta 250 puestos de trabajo |
| sexo | Hombre |
| rango_edad | Mayor o igual a 45 y menor a 50 años de edad |
| rango_salarial | mayor a 2 y hasta 3 veces el salario mínimo |
| rango_uma | mayor a 4 y hasta 5 veces la UMA |
| Tipo | Urbano |
| Tiempo | Permanente |
| Tipo_tiempo | Urbano-Permanente |
| Tarifa_base | 354.8 |
| Asegurados | 2 |
| Masa | 709.6 |

*Fuente: Elaboración propia con el diccionario de datos*

Es importante recalcar que, con esta estructura, se reducen las columnas, ya que no se muestran valores de cero en la cantidad de puestos o en la masa salarial, así que las gráficas y modelos serán más fáciles de construir al haber creado estas variables categóricas.

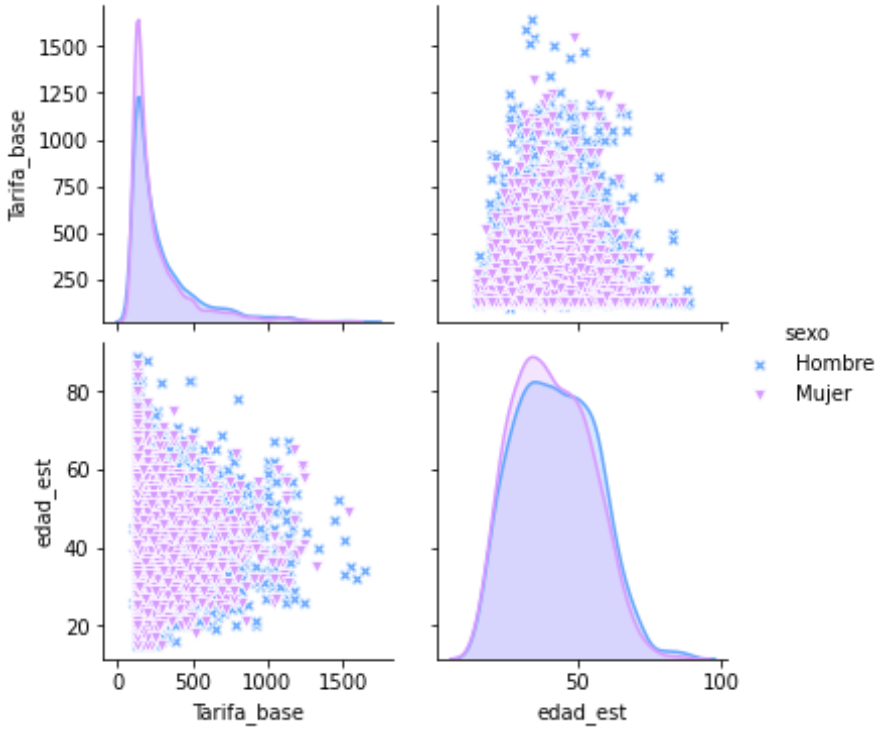
## Propuesta

### Inteligencia de Negocio

Considere el siguiente escenario, se quiere encontrar la relación entre la tarifa base de cotización y la edad, separada por sexo para los puestos afiliados al IMSS en el sector económico denominado “Servicios personales para el hogar y diversos”. La razón es descubrir la existencia o inexistencia de brecha salarial por género, además de demostrar si la edad es una variable definitiva para el crecimiento o decrecimiento de cuánto percibe la persona al día.

Un nuevo desafío es convertir la variable categórica desplegada como un rango de edad a valores numéricos que puedan compararse gráficamente contra la tarifa base de cotización. Una opción sería presentar la cota superior o inferior del rango de edad, que está separado por quinquenios, sin embargo no es correcto asumir que las personas en cuestión tienen la edad “redondeada” a quinquenios. Por ello, la forma empleada para resolverlo es presentar un número aleatorio entre el rango definido en la variable categórica “rango_edad”. La mejora de este método contra el anterior radica en la presentación gráfica.

**Gráfica 4. Brecha salarial vs edad para el sector empleado en el hogar**

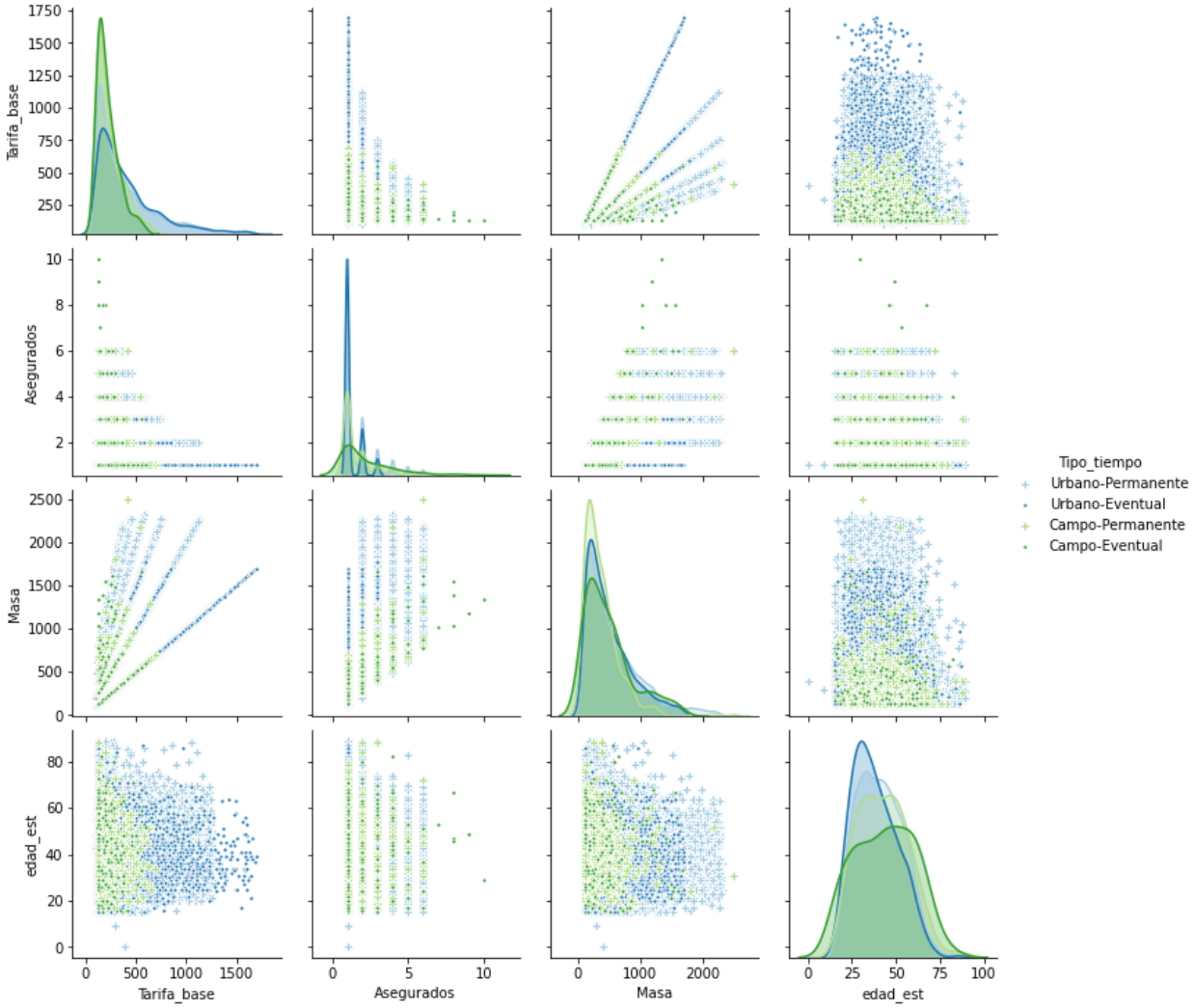



*Fuente: Elaboración propia con el conjunto de datos trabajado*

En efecto, desafortunadamente para las mujeres empleadas en el sector económico del hogar (y afiliadas al seguro social), la tarifa base de cotización que poseen tiene una curtosis más leptocúrtica que sus homólogos masculinos. Además, el conjunto de datos cuenta con registros femeninos más jóvenes afiliados en este sector y no por ello explica la desviación de lo que ganan al día, porque no parece haber una relación positiva entre la edad y la tarifa base de cotización, al menos para este sector tan esencial e infravalorado.

Por otro lado, la brecha salarial no sólo existe comparando géneros, sino en el tipo y la temporalidad del puesto.

**Gráfica 5. Pairplot de variables numéricas en el conjunto de datos**



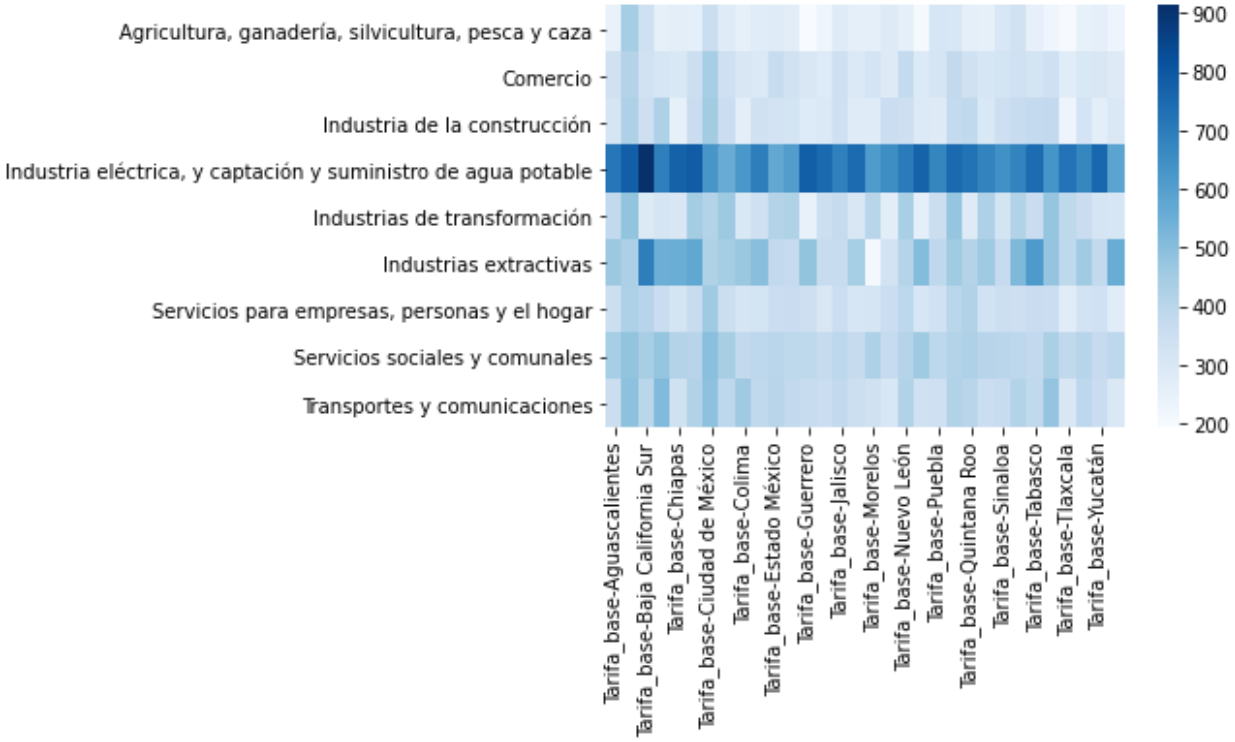

*Fuente: Elaboración propia con el conjunto de datos trabajado*

Gracias a este gráfico, se pueden obtener conclusiones interesantes.

- Los puestos de campo presentan le menor tarifa base de cotización, pero los permanentes presentan mayor masa salarial, significa muchísimos más puestos y mal pagados
- Hay más personas jóvenes laborando en el sector urbano de manera eventual, el sector-temporalidad mejor pagado
- Para el sector urbano entre más puestos permanentes se tengan, no necesariamente se traduce en mucho mejor tarifa base de cotización para sus trabajadores. Es decir, la relación entre masa salarial y tarifa base es positiva pero débil para este tipo de sector-temporalidad.

Ahora, ¿Cómo se comporta la tarifa base de cotización entre estados y sectores?

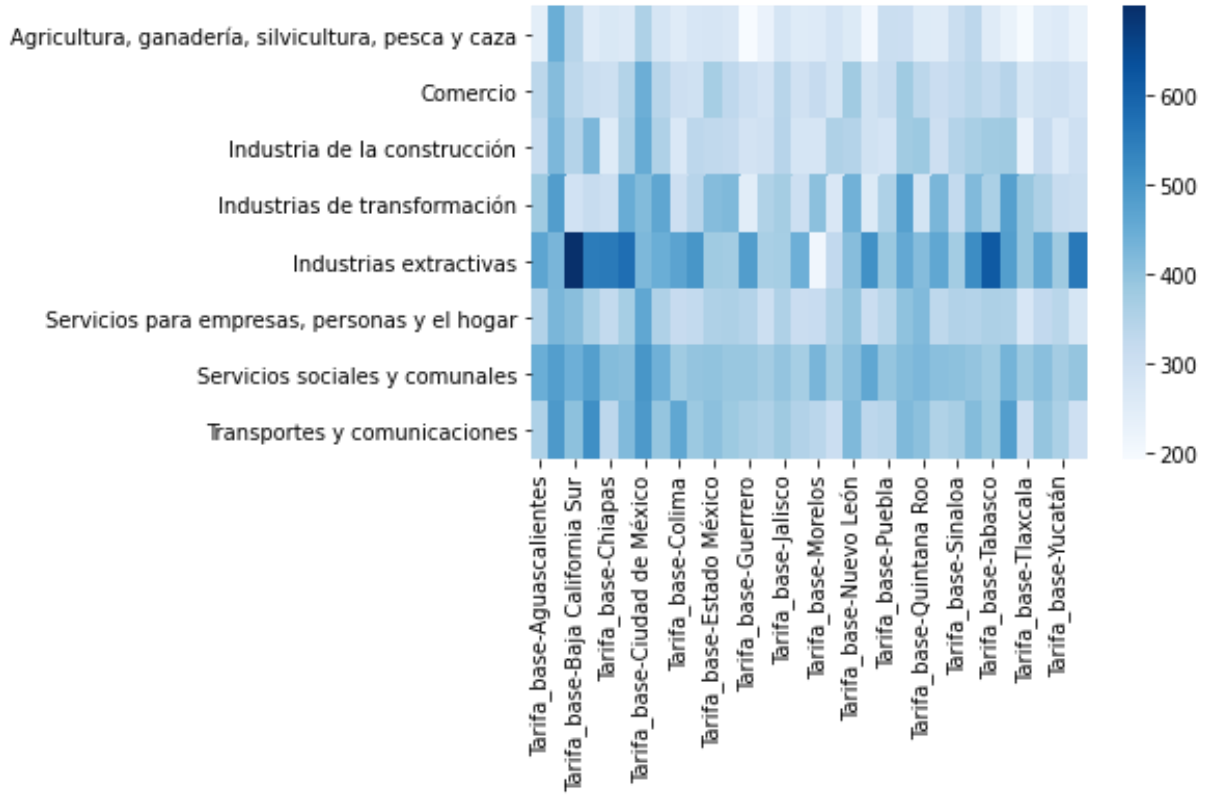
**Gráfica 6. Mapa de calor para la tarifa base de cotización por entidad y sector económico**




*Fuente: Elaboración propia con el conjunto de datos trabajado*

Al parecer, el sector de electricidad y agua tiene una tarifa base de cotización inusualmente alta. Presentando la misma gráfica, pero omitiendo este sector económico, es posible distinguir la diferencia de cómo ganaría la misma persona en el mismo sector, pero en diferente entidad del país, o cómo en un mismo estado dependiendo el sector al que decidas dedicarte definirá radicalmente tus ingresos, tu calidad de vida.

**Gráfica 7. Mapa de calor para la tarifa base de cotización por entidad y sector económico (sin sector económico referente a electricidad y agua)**




*Fuente: Elaboración propia con el conjunto de datos trabajado*

## Analítica Avanzada

Finalmente, como propuesta para un análisis estadístico más avanzado se proponen dos modelos:

- Algoritmo de clústeres por municipios o subdelegaciones de acuerdo con la tarifa base de cotización, sector económico, tipo y temporalidad de puestos para definir estratos económicos en el país

- Modelo de pronóstico para, si deseo poner una empresa de determinado sector (aunado al tipo y temporalidad de los puestos que serán creados) y en cierta entidad/delegación (e incluso conociendo el estrato económico encontrado gracias al inciso anterior), a manera de intervalo conocer el tamaño ideal de mi futura plantilla de trabajadores, así como la tarifa base de cotización digna para cada empleado/a, tal vez tomando en cuenta la edad, pero definitivamente no diferenciar por el género de cada persona.

Sí, con la suficiente cantidad de registros de calidad, una tabla de caracteres alfanuméricos puede ser organizada y segmentada para encontrar patrones y tomar decisiones que nos lleven a todos hacia un mundo mejor.