UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

MAESTRÍA EN CIENCIA DE DATOS

**APRENDIZAJE AUTOMÁTICO**

Profesor: Jose Anastacio Hernandez Saldaña

**Tarea 1:** Análisis Exploratorio

Dora Alicia Guevara Villalpando

Matrícula: 1551003

Grupo: 003

30 de junio de 2024

**Tarea 1.** Análisis Exploratorio

# Introducción.

El presente documento presenta los resultados obtenidos durante el proceso de realizar la Tarea 1 de Aprendizaje Automático. Esta tarea consiste en realizar el análisis exploratorio a una base de datos proporcionado en clase, esta base de datos cuenta con 5 columnas y 636,201 filas.

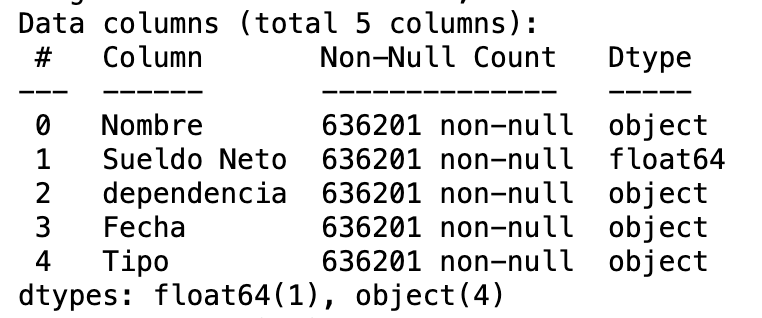
# Desarrollo.

Durante la realización de esta tarea se llevaron a cabo cinco actividades principales:

1. Identificar las entidades más representativas.
2. Obtener estadísticas descriptiva de cada entidad.
3. Hacer agrupaciones por las entidades y sacar estadisticas de las agrupaciones.
4. Crear imagenes de estas estadisticas, ya sean histogramas, graficas de pastel, etc.
5. De alguna de las agrupaciones, hacer una pureba ANOVA.

A continuación se describen los hallazgos en cada una de las actividades realizadas.

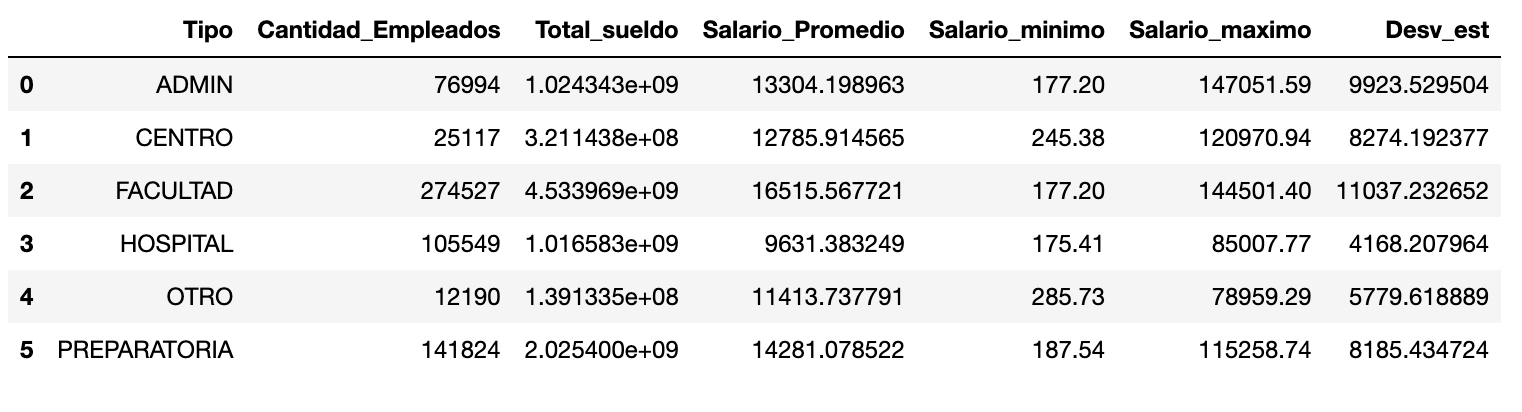
Nuestra base de datos esta conformada por 5 columnas (variables), las cuales representan diferentes tipos de datos a analizar:



En esta base de datos contamos con variable del tipo numérico (float): *Sueldo Neto*, y las demás estan clasificadas “objetos” de caracteres: *Nombre* (nombre del empleado), *dependencia* (nombre de la institución registrada en la UANL), *Fecha* y *Tipo* (clasificación de las dependencias registradas en administración, centro, preparatoria, facultad, hospital y otro).

Previo a iniciar con los análisis de estadística descriptiva decidí buscar si en las variables de la base de datos hay valores nulos para poder identificarlos y decidir que hacer con ellos; sin embargo, esta busqueda dio como resultado que en la base de datos no hay valores nulos por lo que procedí a obtener los estadísticos descriptivos.

**Análisis descriptivos.**



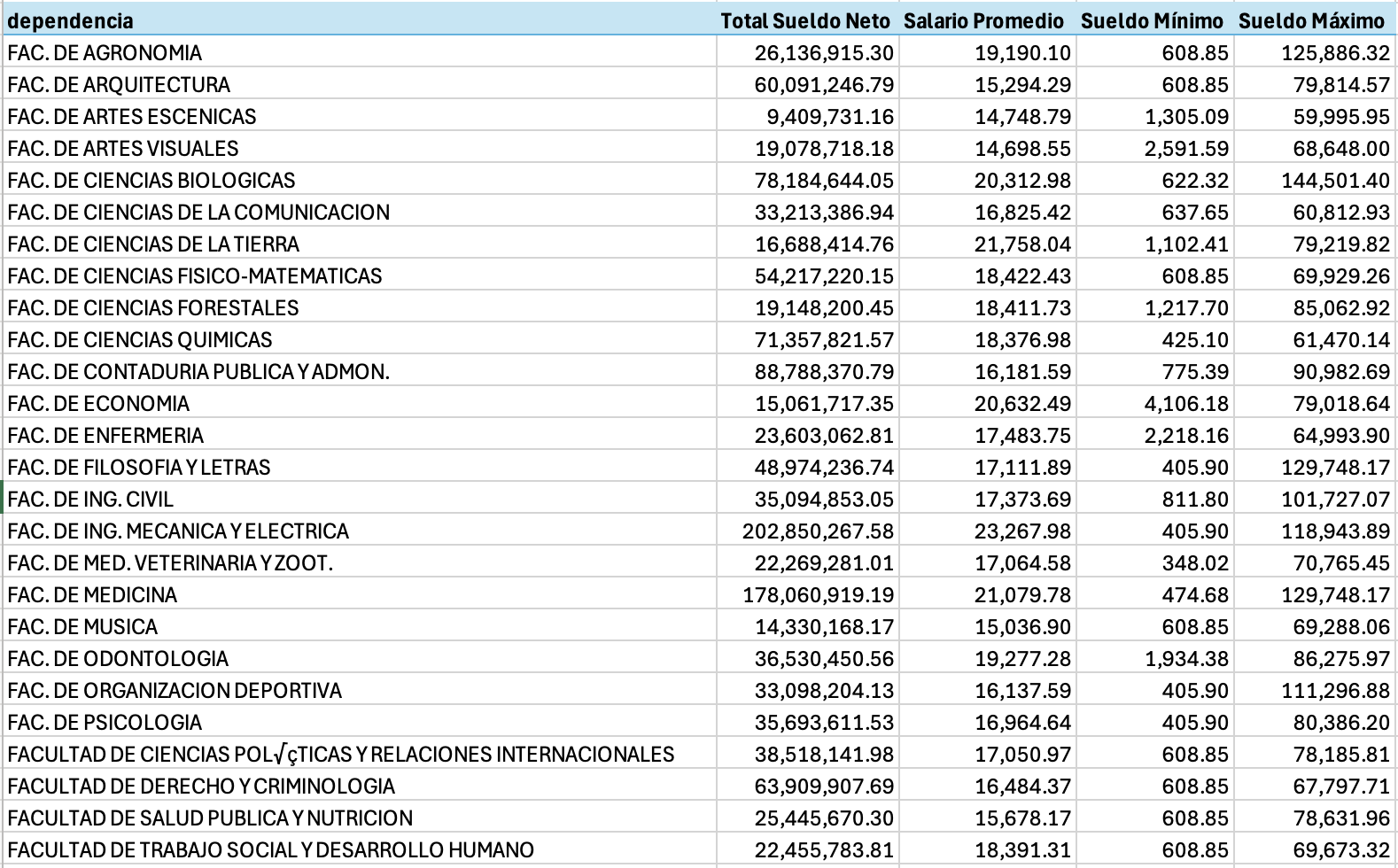
Se analizaron los estadísticos descriptivos de las dependencias utilizando la categría pre-establecida en la columna *Tipo*. En estos resultados se observa que el sueldo acumulado (total\_sueldo) máximo se encuentra en el Tipo = FACULTAD, así como el salario promedio más alto de 16,515.57.

Se decició obtener la desviación estándar de cada *Tipo* ya que nos sirve para hacer una estimación sobre cómo de dispersos están los datos con respecto a la media. Se observa que todas las desviaciones estándar son altas y esto representa mayor dispersión de los datos conforme a la media; esto se puede concluir al observar el rango tan amplio que hay entre el Salario mínimo y el Salario máximo.

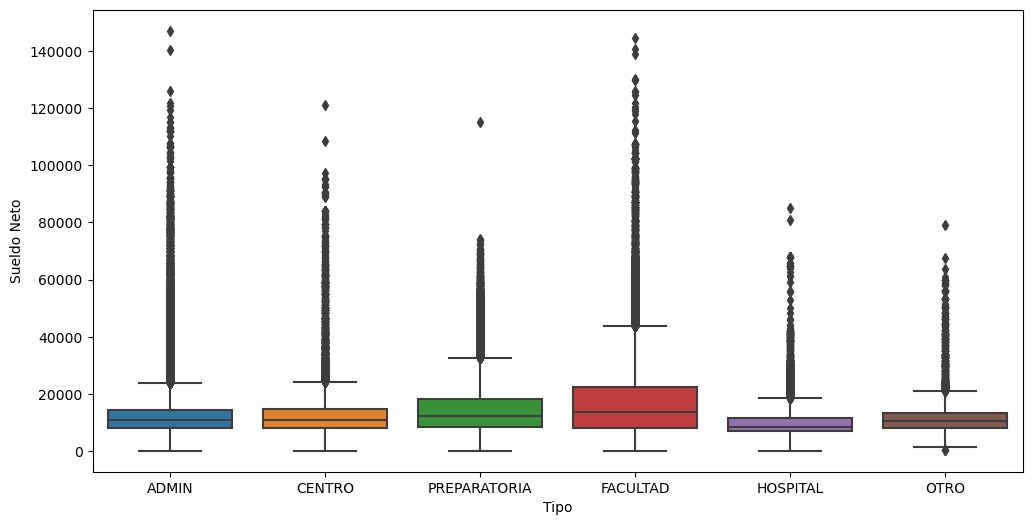
También se hizo un análisis similar al de la tabla anterior pero agregando el año, esto para identificar patrones o diferencias. Se observó que en 2019 se tiene una cantidad menor de empleados registrados y por ende un sueldo acumulado menor a los demás años. La excepción sería la información que se muestra del 2024, ya que al ser el año corriendo la información no está completa.

Con las estadísticas descriptivas se identifica que la FACULTAD durante el 2023 es donde se tiene la mayor cantidad de empleados y se maneja el salario máximo. Se observó lo siguiente:

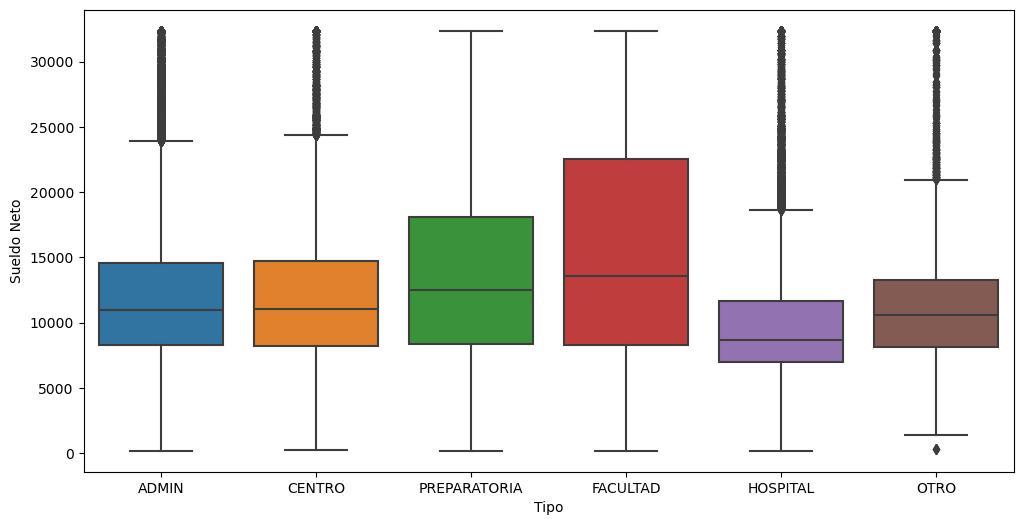
* La facultad que cuenta con el salario promedio más alto es *FIME*.
* La facultad que cuenta con el salario máximo es *Ciencias Biologicas*.
* La facultad de *Artes Visuales* cuenta con el salario promedio más bajo de todas.
* La facultad de *Artes Escenicas* cuenta con el salario máximo más pequeño de todas las facultades aunque su salario promedio es mayor al de la facultad de artes visuales.



Se realiza una gráfica de cajas para cada uno de los *Tipos* de dependencias:



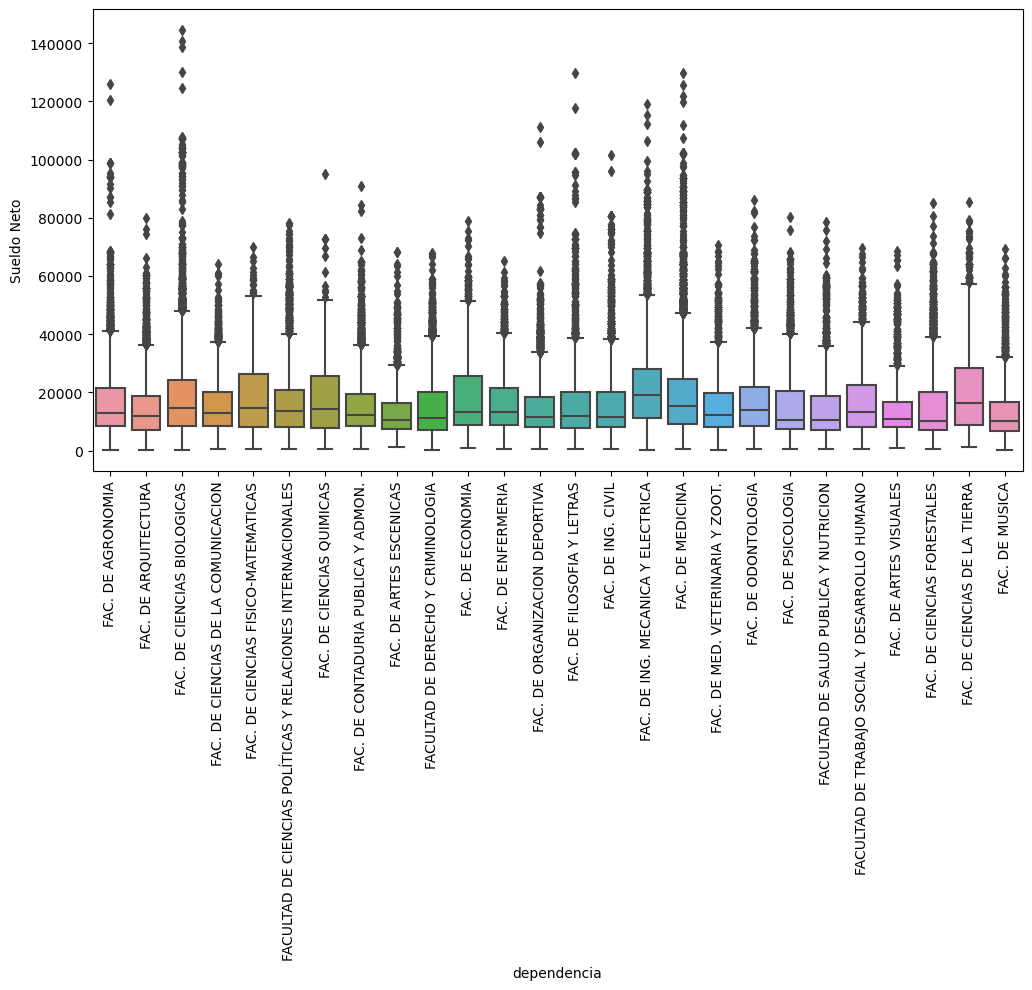
Reducimos los outliers para poder apreciar el diagrama de caja de una mejor forma:



Se concluye lo siguiente:

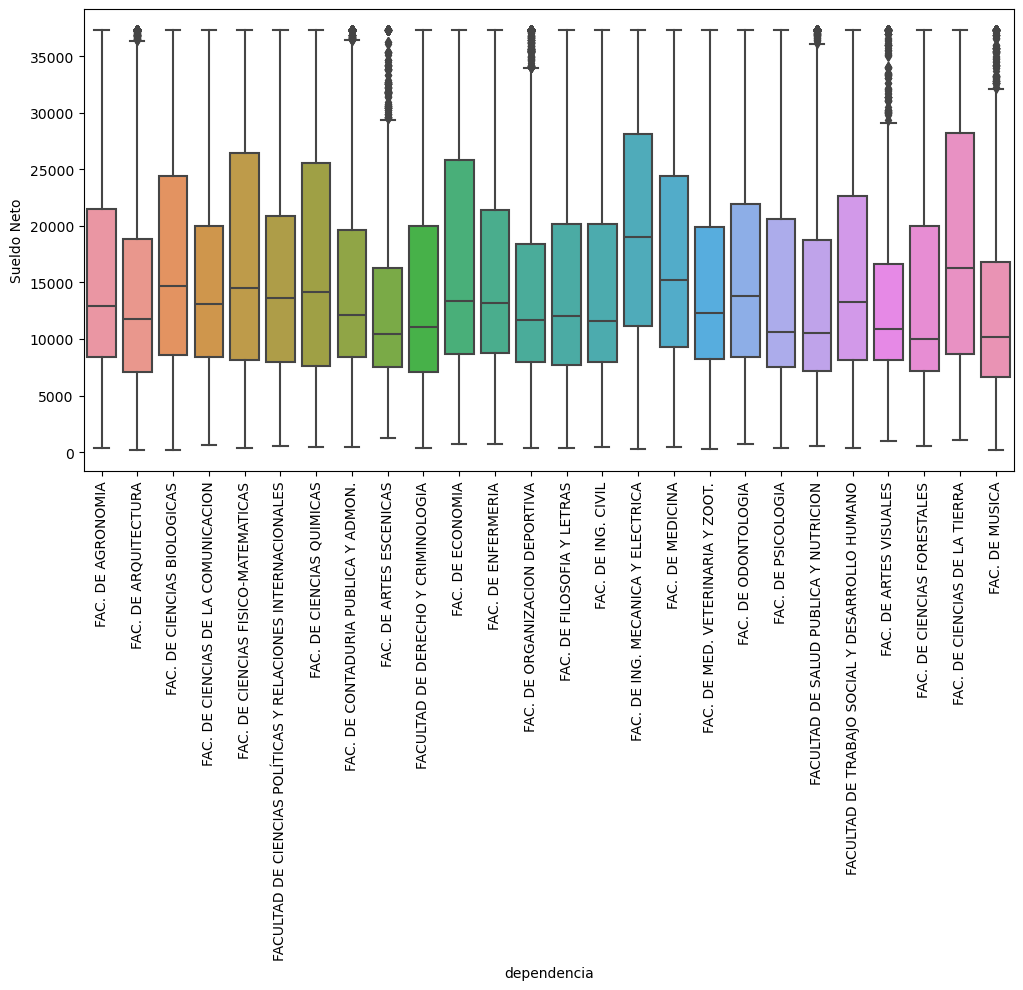
* La caja de FACULTAD es más amplia que las demás.
* Todas tienen valores atípicos (outliers).
* La caja que parece mostrar menor dispersión en los datos alrededor de la media es OTRO.
* Se observa que el salario promedio de ADMIN y CENTRO es similar.
* El salario promedio más alto se encuentra en FACULTAD,
* El salario promedio más pequeño se encuentra en HOSPITAL.

A continuación se presenta un conjunto de gráficos de caja para cada una de las facultades:



La facultad de *ciencias biologicas* cuentan con un rango más disperso de *sueldo neto*.

Reducimos outliers para observar mejor la información:

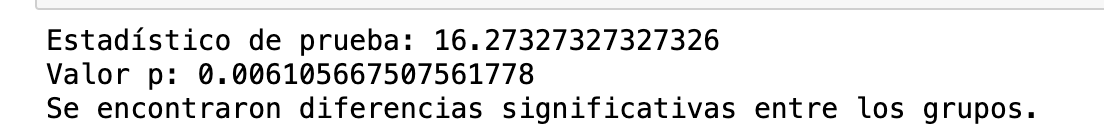


* FIME cuenta con el salario promedio más alto.
* La segunda facultad con el salario promedio más alto es MEDICINA.

**ANOVA:**

No hay diferencias entre las medias de los diferentes grupos.

Al menos un par de medias son significativamente distintas la una de la otra.



Como entonces rechazamos

Se infiere que al menos un par de medias son signifivativas por lo que no hay normalidad.

