# Proyecto Final: Procesamiento de Datos - Grupo N°18

En la recolección se hace una carga del data set en la notebook que contiene un archivo csv listo para empezar el desarrollo.

Analizamos las dimensiones del data set en el dataframe y Como resultado tenemos **3755 filas y 11 columnas**

Analizando el data set podemos concluir que tenemos lo siguiente:
1-El data set tiene 3755 filas y 11 columnas.
2-Contiene datos del tipo string(object) e int(enteros).
3-No tiene faltantes de datos.

Limpieza

Columnas irrelevantes: observamos que las columnas "salary" y "salary_currency" no son necesarias para el analisis, ya que la columna salary_in_usd se encarga de representar el salario de los empleados de forma comparativa.
Cambio de nombres en atributos: para un mejor analisis se procede a cambiar el nombre de los atributos para traducirlos de ingles a español, con el fin de que se adapte al lector objetivo.

Registros repetidos: se debe verificar que no se repitan registros en el data set, porque podrian perjudicar el resultado del análisis.

Valores extremos: nos permite ver si hay datos que no son adecuados, valores fuera de un rango, etc.
"IQR method" para detectar Outliers rango intercuartil, como medida de dispersion
Q1: 84975.0 90025.0
Q2: 130000.0
Q3: 175000.0

valores normales o atípicos

limite superior leve: 310037.5 limite inferior leve: -50062.5
limite superior extremo: 445075.0 limite inferior extremo: -185100.0

Conclusión Valores extremos: No se detectan valores negativos en salarios, y dado que la variable analizada puede estar relacionada a otras variables como el título, nivel de experiencia , es que NO excluiremos estos valores atípicos.

Resumen de datos estadísticos que forman parte del conjunto de datos.
Observación:
Podemos observar que hay una cantidad de 2584 registros lo que indica que no hay valores nulos.
Presenta un promedio de 133409 en el salario en usd.
Tiene un desvio estandar minimo de 0.75 en el año de trabajo.
El valor minimo mas bajo es de 0 que representa la presencialidad en el ratio_remoto y el valor maximo es de 450000 que representa el salario.

Observaciones de Histogramas de variables numéricas

La mayor parte del grupo de salarios tiene entre los 50000 y los 200000 usd, con sesgo hacia los 10000-150000 usd.
La mayor parte del grupo de las años de trabajo (75%) tiene un mayor incremento superando por poco a el año 2022.

Nivel_de_experiencia
Senior 1554
Mid/Intermediate level 664
Entry level 270
Executive level 96

Tipo_de_empleo
Full-time 2547
Part-time 17
Contractor 10
Freelancer 10

Cantidades por Residencia de empleado
Residencia_empleado
USA 1893
GBR 150
CAN 81
IND 70
ESP 47
...
BIH 1
ARM 1
CYP 1
KWT 1
MLT 1

Interpretación

La correlación entre 'Año_de_trabajo' y 'Año_de_trabajo' es 1, ya que es la misma variable. Esto indica una correlación perfecta.
La correlación entre 'Año_de_trabajo' y 'Salario_en_usd' es 0.24. Esto sugiere una correlación positiva débil entre el año de trabajo y el salario en dólares.
La correlación entre 'Año_de_trabajo' y 'Ratio_remoto' es -0.22. Esto indica una correlación negativa moderada entre el año de trabajo y el ratio remoto.
La correlación entre 'Salario_en_usd' y 'Salario_en_usd' es 1, ya que es la misma variable. Esto también indica una correlación perfecta.
La correlación entre 'Salario_en_usd' y 'Ratio_remoto' es -0.085. Esto sugiere una correlación negativa muy débil entre el salario en dólares y el ratio remoto.
La correlación entre 'Ratio_remoto' y 'Ratio_remoto' es 1, ya que es la misma variable. Esto indica una correlación perfecta

# Como analisis final de la exploracion realizada sobre los datos podemos concluir en las siguientes conclusiones:

Podemos observar que la mayor concentración de salario se encuentra por los valores minimos en el 2020 pero a medida que avanza los años se fue moviendo y creciendo mas para la derecha a causa de la gran cantidad de demanda de empleos en esta area.

- La mayoria de los puestos ocupados son en niveles Senior, e intermedios en menor medida
- No hay grandes diferencias en el salario entre los trabajos remotos y los "on site"
- La gran mayoria de empleos son full time, por amplia diferencia
- El promedio general de salarios en el mundo de la ciencia de datos viene en aumento