Desafío - Estadística descriptiva y Probabilidades Parte I - Freddy González

```
import pandas as pd
import numpy as np
```

Importar Datos csv

```
df = pd.read_csv('/content/ds_salaries.csv')
print(df)
```

```
      work_year experience_level employment_type                 job_title
0          2023               SE              FT  Principal Data Scientist
1          2023               MI              CT               ML Engineer
2          2023               MI              CT               ML Engineer
3          2023               SE              FT            Data Scientist
4          2023               SE              FT            Data Scientist
...         ...              ...             ...                       ...
3750       2020               SE              FT            Data Scientist
3751       2021               MI              FT  Principal Data Scientist
3752       2020               EN              FT            Data Scientist
3753       2020               EN              CT     Business Data Analyst
3754       2021               SE              FT      Data Science Manager

       salary salary_currency  salary_in_usd employee_residence  remote_ra
0       80000             EUR          85847                ES
1       30000             USD          30000                US
2       25500             USD          25500                US
3      175000             USD         175000                CA
4      120000             USD         120000                CA
...       ...             ...            ...               ...
3750   412000             USD         412000                US
3751   151000             USD         151000                US
3752   105000             USD         105000                US
3753   100000             USD         100000                US
3754  7000000             INR          94665                IN

     company_location company_size
0                  ES            L
1                  US            S
2                  US            S
3                  CA            M
4                  CA            M
...               ...          ...
3750               US            L
3751               US            L
3752               US            S
3753               US            L
3754               IN            L

[3755 rows x 11 columns]
```

1. Analicemos el promedio, desviación estándar, quintiles y rango para la columna salary_in_usd.

```
# Cálculos estadísticos básicos
mean_salary = df['salary_in_usd'].mean()
std_salary = df['salary_in_usd'].std()
quantiles = df['salary_in_usd'].quantile([0.25, 0.5, 0.75])
salary_range = df['salary_in_usd'].max() - df['salary_in_usd'].min()

# Mostrar los resultados
print(f"Promedio: {mean_salary}")
print(f"Desviación estándar: {std_salary}")
print(f"Quintiles:\n{quantiles}")
print(f"Rango: {salary_range}")
```

```
Promedio: 137570.38988015978
Desviación estándar: 63055.625278224084
Quintiles:
0.25     95000.0
0.50    135000.0
0.75    175000.0
Name: salary_in_usd, dtype: float64
Rango: 444868
```

2. Comparación por categorías elejimos 3 categorías: experience_level, employment_type, company_size

```
categories = ['experience_level', 'employment_type', 'company_size']

for category in categories:
    print(f"Análisis por {category}:")
    print(df.groupby(category)['salary_in_usd'].agg(['mean', 'median', 'st
    print("\n")
```

```
Análisis por experience_level:
                          mean    median           std    min     max
experience_level
EN                 78546.284375   70000.0  52225.424309   5409  300000
EX                194930.929825  196000.0  70661.929661  15000  416000
MI                104525.939130  100000.0  54387.685128   5132  450000
SE                153051.071542  146000.0  56896.263954   8000  423834


Análisis por employment_type:
                          mean    median           std    min     max
employment_type
CT                113446.900000   75000.0  130176.746842   7500  416000
FL                 51807.800000   50000.0   29458.879336  12000  100000
FT                138314.199570  135000.0   62452.177613   5132  450000
PT                 39533.705882   21669.0   38312.145181   5409  125404


Análisis por company_size:
                       mean    median           std   min     max
company_size
L             118300.982379  108500.0  75832.391505  5409  423834
M             143130.548367  140000.0  58992.813382  5132  450000
S              78226.682432   62146.0  61955.141792  5679  416000
```

Más representativa: experience_level los niveles de experiencia tienen menor dispersión en los salarios y un patrón más claro. Menos representativa: employment_type existe alta variabilidad entre tipos de empleo (por ejemplo, entre freelance y tiempo completo), lo que afecta la representatividad de las medidas centrales.

3. Análisis de cargos en empresas de EEUU

```
#Filtraremos las empresas con sede en EE.UU. (company_location == 'US') y
# Filtrar empresas con sede en EE.UU.
usa_jobs = df[df['company_location'] == 'US']

# Analizar los mejores salarios por cargo
top_jobs_in_usa = usa_jobs.groupby('job_title')['salary_in_usd'].mean().so

# Mostrar los resultados
print("Cargos mejor pagados en empresas de EE.UU.:")
print(top_jobs_in_usa)
```

```
Cargos mejor pagados en empresas de EE.UU.:
job_title
Data Analytics Lead                   405000.000000
Data Science Tech Lead                375000.000000
Director of Data Science              294375.000000
Principal Data Scientist              255500.000000
Cloud Data Architect                  250000.000000
Applied Data Scientist                238000.000000
Head of Data                          233183.333333
Machine Learning Software Engineer    217400.000000
Data Lead                             212500.000000
Head of Data Science                  202355.000000
Name: salary_in_usd, dtype: float64
```