**ANALYSIS OF DATASET FROM KAGGLE TO MAKE THE DATASET MORE USEABLE**

# Which dataset you've selected?

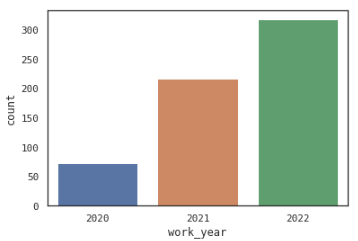
The data set that I have selected is Data Science Job Salaries containing salaries of jobs in the data science domain containing the following tables.

|  |  |
| --- | --- |
| work\_year | The year the salary was paid. |
| experience\_level | The experience level in the job during the year with the following possible values: EN Entry-level / Junior MI Mid-level / Intermediate SE Senior-level / Expert EX Executive-level / Director |
| employment\_type | The type of employement for the role: PT Part-time FT Full-time CT Contract FL Freelance |
| job\_title | The role worked in during the year. |
| salary | The total gross salary amount paid. |
| salary\_currency | The currency of the salary paid as an ISO 4217 currency code. |
| salary*in*usd | The salary in USD (FX rate divided by avg. USD rate for the respective year via fxdata.foorilla.com). |
| employee\_residence | Employee's primary country of residence in during the work year as an ISO 3166 country code. |
| remote\_ratio | The overall amount of work done remotely, possible values are as follows: 0 No remote work (less than 20%) 50 Partially remote 100 Fully remote (more than 80%) |
| company\_location | The country of the employer's main office or contracting branch as an ISO 3166 country code. |
| company\_size | The average number of people that worked for the company during the year: S less than 50 employees (small) M 50 to 250 employees (medium) L more than 250 employees (large) |

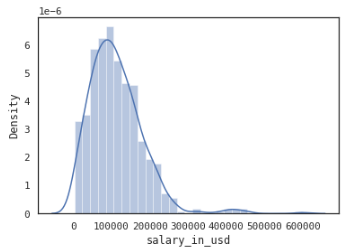
# 2) What analysis you've done in the starter code?

I started by importing all the libraries the I displayed the salary then I deleted some rows by using the drop method of python which were unnamed i.e.

Then I displayed the sum, mean max min of the salaries all the info of the salaries then displayed who had the minimum and who had the maximum salary, then I displayed the salaries in a graphical way



Grouped up the salaries into chunks of same salaries took out the average salary in USD and another graph of work year and salaries in USD displayed a histogram of the salary in USD



Displayed all the job title in the csv file and the count of the all the employees with their job title and we grouped that by their job title and the location of their company displayed their experience level a graph showing their experience level and another graph showing their average salary by experience level the size of the company and the least salary of data scientist by country.

# What information you got?

The information that I got from the dataset and the python code is that It classifies the data into sets and groups which can be used then rather than using the huge data set which is incontinent and this classification makes the he data set in more functional form.