Documentation Flow

Project - 2

EDA-1

A) Problem Statement->

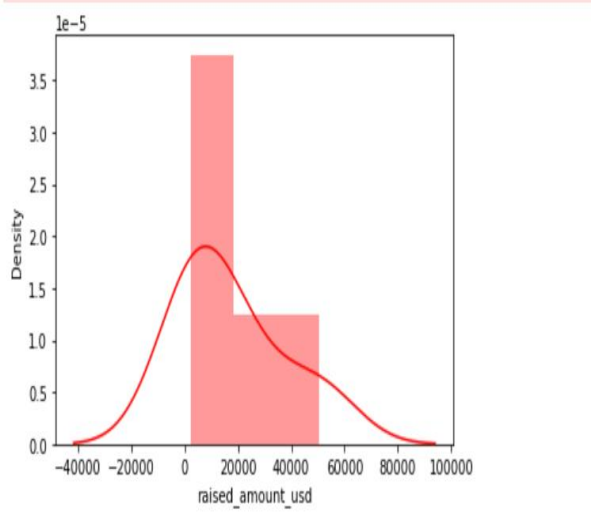
We are provided with real time companies dataset and we have to perform EDA on that to find out solution to some needed excel questions which would be beneficial for the company to make some decisions in future.

B)Overview ->

Firstly we did the analysis of the data and then checked for null values and cleared them, then we found out the solution to the questions using different command and then later we made some plots to visualize the data in a better way.

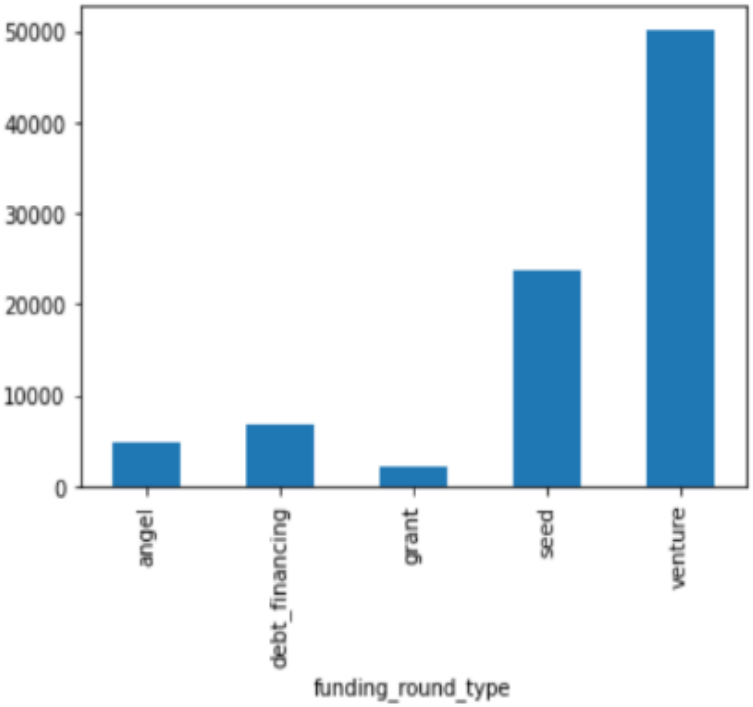
C)Step-wise (plots, inferences, preprocessing)->

Step-1->



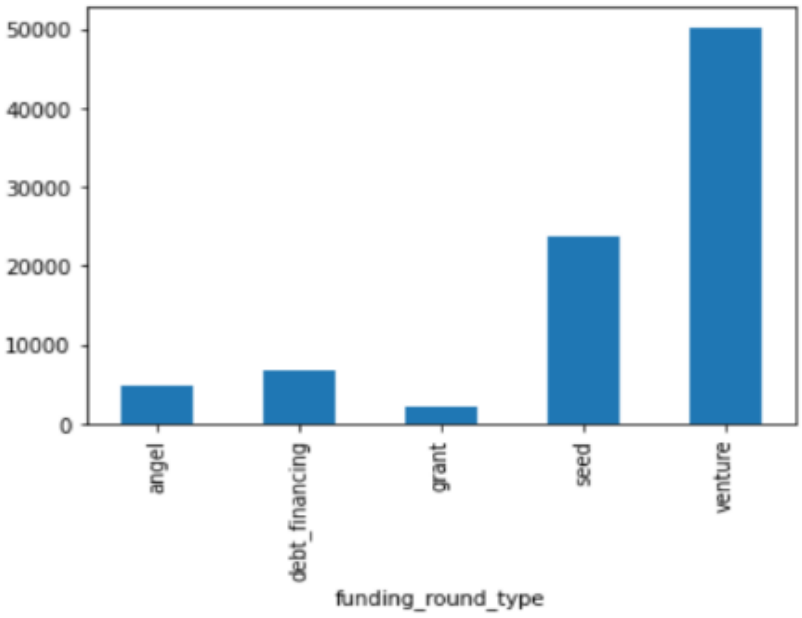
Inference from this -> how is our plot distributed, kde depicts the probability density function of the continuous or non-parametric data variables

Step-2->



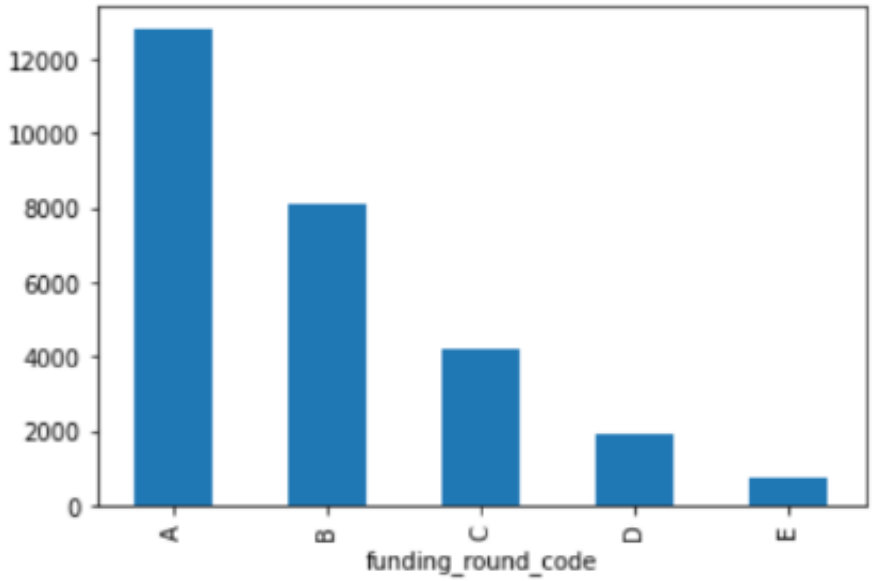
Inference from this -> how is funding round type related with raised amount usd.

Step-3->



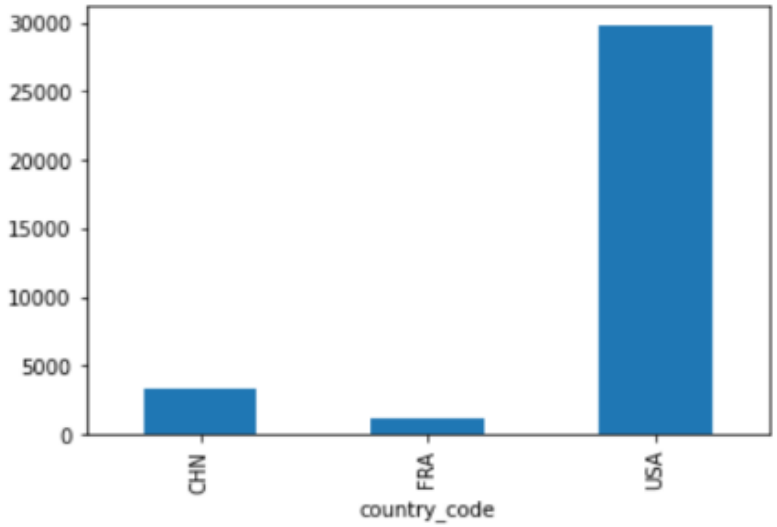
Inference from this -> Mean values plot between round type related with raised amount usd.

Step-4->



Inference from this -> how is funding round code related with raised amount usd.

Step-5->



Inference from this -> how is country code related with raised amount usd.

A) Jump to the questions {High level view for each exercise}-

Table-1: Questions

1	How many unique companies are present in rounds2?
2	How many unique companies are present in the companies file?
3	In the ***companies*** data frame, which column can be used as the unique key for each company? Write the **name of the column**.
4	Are there any companies in the rounds2 file which are not present in ***companies*** ? Answer **Y/N**.
5	Merge the two data frames so that all variables (columns) in the ***companies*** frame are added to the ***rounds2*** data frame. Name the merged frame ***master_frame***. How many observations are present in ***master_frame ?***

Table-1: Answers

66373
66368
permalink i.e. Company_Link
Y
Name is master_frame2 and number of observations are 114943

Table-2 : Questions

Sl.No	Questions
1	Representative funding amount of venture type
2	Representative funding amount of angel type
3	Representative funding amount of seed type
4	Representative funding amount of private equity type
5	Considering that Spark Funds wants to invest between **5 to 15 million USD** per investment round, which investment type is the most suitable for them?

Table-2 : Answers

Answer
1.17E+07
9.59E+05
7.20E+05
7.33E+07
5.0E+06 to 1.5E+07 venture

Table-3 : Questions

Sl.No	Questions
1	Top English speaking country
2	Second English speaking country
3	Third English speaking country

Table-3 : Answers

Answers
USA
GRB
JPN

Table-4 : Questions

Sl.no	Questions
1	Total number of Investments (count)
2	Total amount of investment (USD)
3	Top Sector name (no. of investment-wise)
4	Second Sector name (no. of investment-wise)
5	Number of investments in top sector (3)
6	Number of investments in second sector (4)

Table-4 : Answers

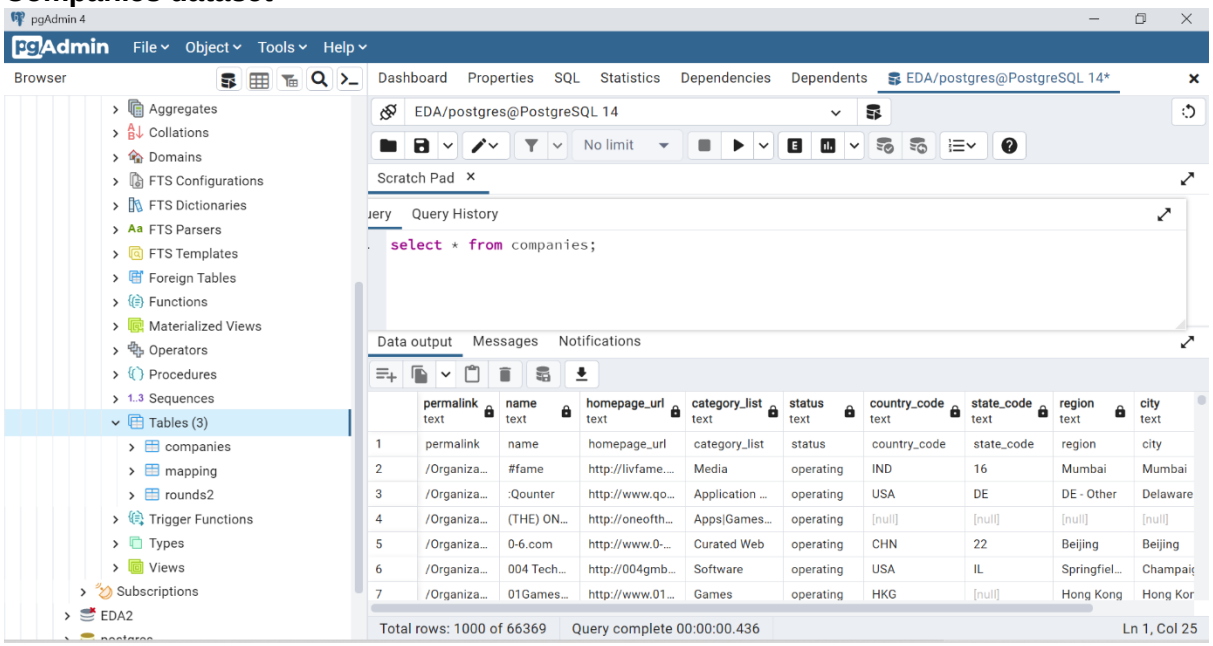
C1	C2	C3
29799	1201	3294
2.72E+11	6.32E+10	5.46E+10
Manufacturing	Entertainment	Manufacturing
Entertainment	Manufacturing	Entertainment
19920	801	2760
9879	400	534

In Table-4 we assume let C1, C2 and C3 according to the top highest raised amount countries, that are:

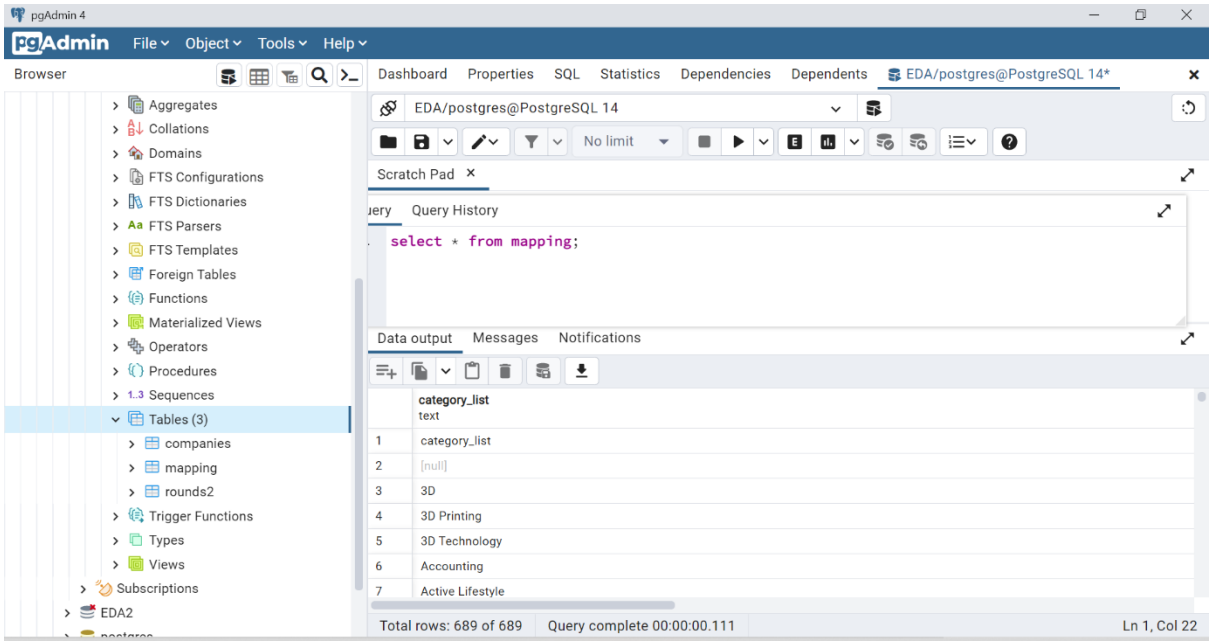
C1 = USA
C2 = FRA
C3 = CHN

Connection with progress-> Fetching the dataset in PgAdmin by creating tables and importing csv files into that .

Companies dataset



Mapping dataset



Rounds dataset