

# **Documentation Flow**

## **Project - 2**

### **EDA-1**

#### **A) Problem Statement->**

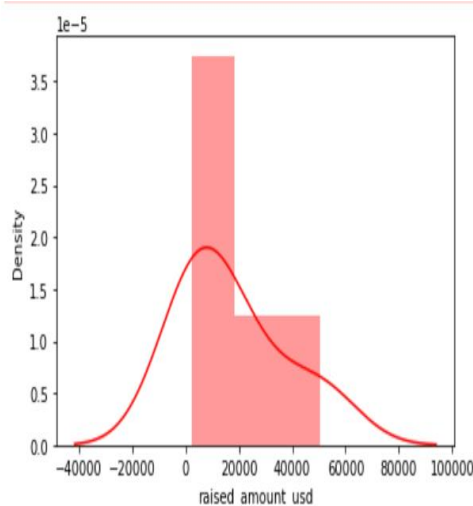
We are provided with real time companies dataset and we have to perform EDA on that to find out solution to some needed excel questions which would be beneficial for the company to make some decisions in future.

#### **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.

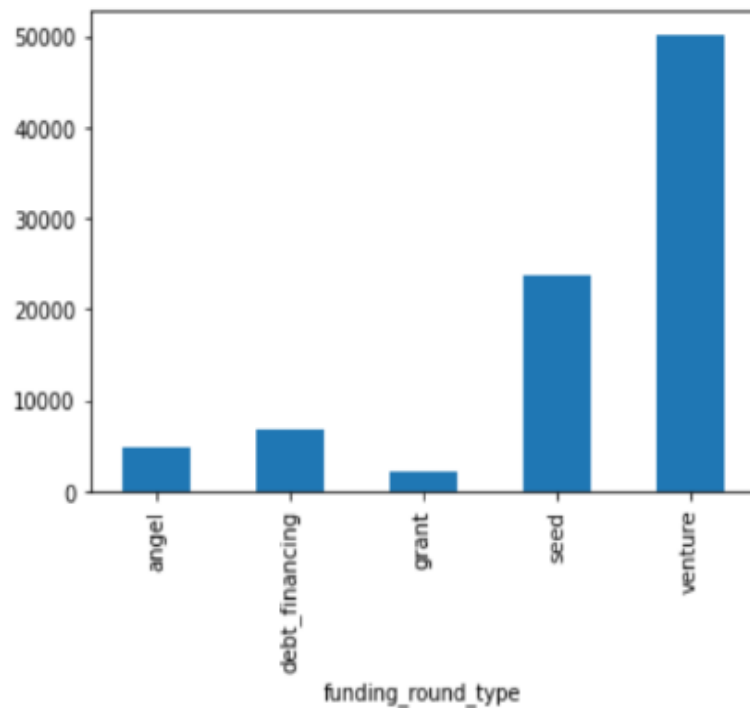
#### **C)Step-wise (plots, inferences, preprocessing)->**

##### **Step-1->**



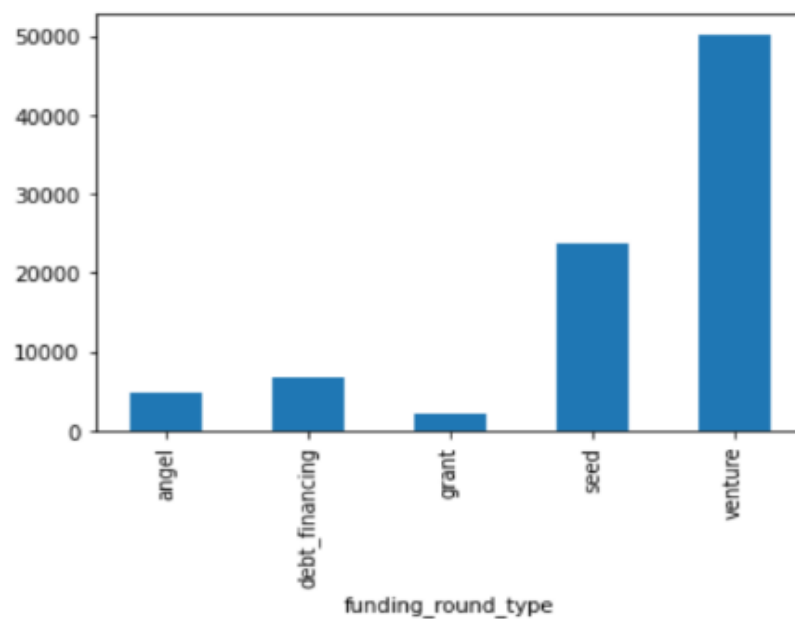
Inference from this -> how is our plot distributed, kde depicts the probability density function of the continuous or non-parametric data variables

**Step-2->**



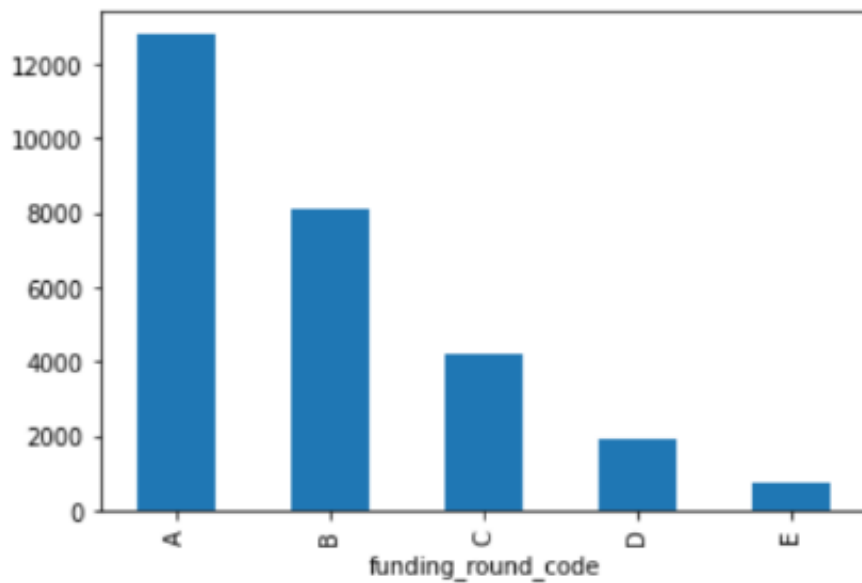
**Inference from this -> how is funding round type related with raised amount usd.**

**Step-3->**



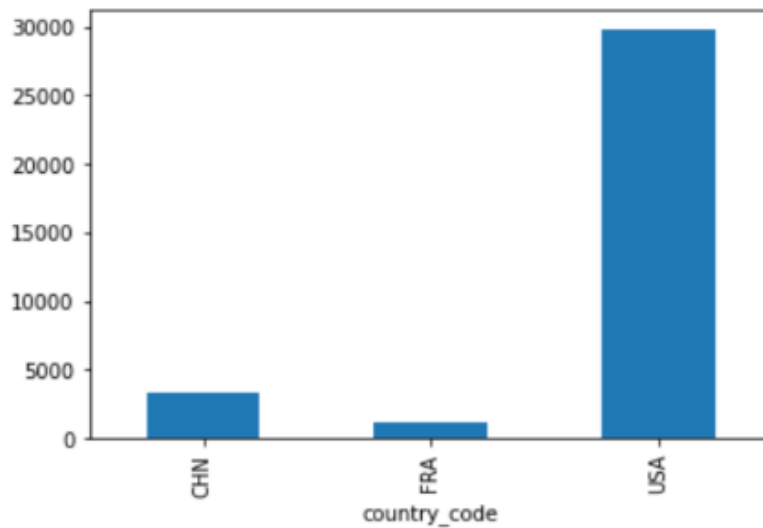
**Inference from this -> Mean values plot between round type related with raised amount usd.**

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

**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 <b>5 to 15 million USD</b> 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

pgAdmin 4

File Object Tools Help

Browser

- Aggregates
- Collations
- Domains
- FTS Configurations
- FTS Dictionaries
- FTS Parsers
- FTS Templates
- Foreign Tables
- Functions
- Materialized Views
- Operators
- Procedures
- Sequences
- Tables (3)**
  - companies
  - mapping
  - rounds2
- Trigger Functions
- Types
- Views
- Subscriptions
- EDA2

Dashboard Properties SQL Statistics Dependencies Dependents EDA/postgres@PostgreSQL 14\*

EDA/postgres@PostgreSQL 14

Scratch Pad x

Query History

```
select * from companies;
```

Data output Messages Notifications

	permalink text	name text	homepage_url text	category_list text	status text	country_code text	state_code text	region text	city text
1	permalink	name	homepage_url	category_list	status	country_code	state_code	region	city
2	/Organiza...	#fame	http://livfame...	Media	operating	IND	16	Mumbai	Mumbai
3	/Organiza...	:Qounter	http://www.qo...	Application ...	operating	USA	DE	DE - Other	Delaware
4	/Organiza...	(THE) ON...	http://oneofth...	Apps/Games...	operating	[null]	[null]	[null]	[null]
5	/Organiza...	0-6.com	http://www.0-...	Curated Web	operating	CHN	22	Beijing	Beijing
6	/Organiza...	004 Tech...	http://004gmb...	Software	operating	USA	IL	Springfiel...	Champaig
7	/Organiza...	01Games...	http://www.01...	Games	operating	HKG	[null]	Hong Kong	Hong Kor

Total rows: 1000 of 66369 Query complete 00:00:00.436 Ln 1, Col 25

## Mapping dataset



pgAdmin 4

File Object Tools Help

EDA/postgres@PostgreSQL 14\*

Scratch Pad

```
select * from mapping;
```

Data output

	category_list text
1	category_list
2	[null]
3	3D
4	3D Printing
5	3D Technology
6	Accounting
7	Active Lifestyle

Total rows: 689 of 689 Query complete 00:00:00.111 Ln 1, Col 22

## Rounds dataset

pgAdmin 4

File Object Tools Help

EDA/postgres@PostgreSQL 14\*

Scratch Pad

```
select * from rounds2;
```

Data output

	company_permalink text	funding_round_permalink text	funding_round_type text	funding_round_code text	funded_at text	raised_amount_usd text
1	company_permalink	funding_round_permalink	funding_round_type	funding_round_code	funded_at	raised_amount_usd
2	/organization/-fame	/funding-round/9a01d05...	venture	B	05-01-2015	10000000
3	/ORGANIZATION/-Q...	/funding-round/22dacff4...	venture	A	14-10-2014	[null]
4	/organization/-qoun...	/funding-round/b44fbb9...	seed	[null]	01-03-2014	700000
5	/ORGANIZATION/-T...	/funding-round/650b8f7...	venture	B	30-01-2014	3406878
6	/organization/0-6-c...	/funding-round/5727acc...	venture	A	19-03-2008	2000000
7	/ORGANIZATION/00...	/funding-round/1278dd4...	venture			

✓ Successfully run. Total query runtime: 156 msec. 114950 rows

Total rows: 1000 of 114950 Query complete 00:00:00.156 Ln 1, Col 22