INVESTMENT ASSIGNMENT

SUBMISSION

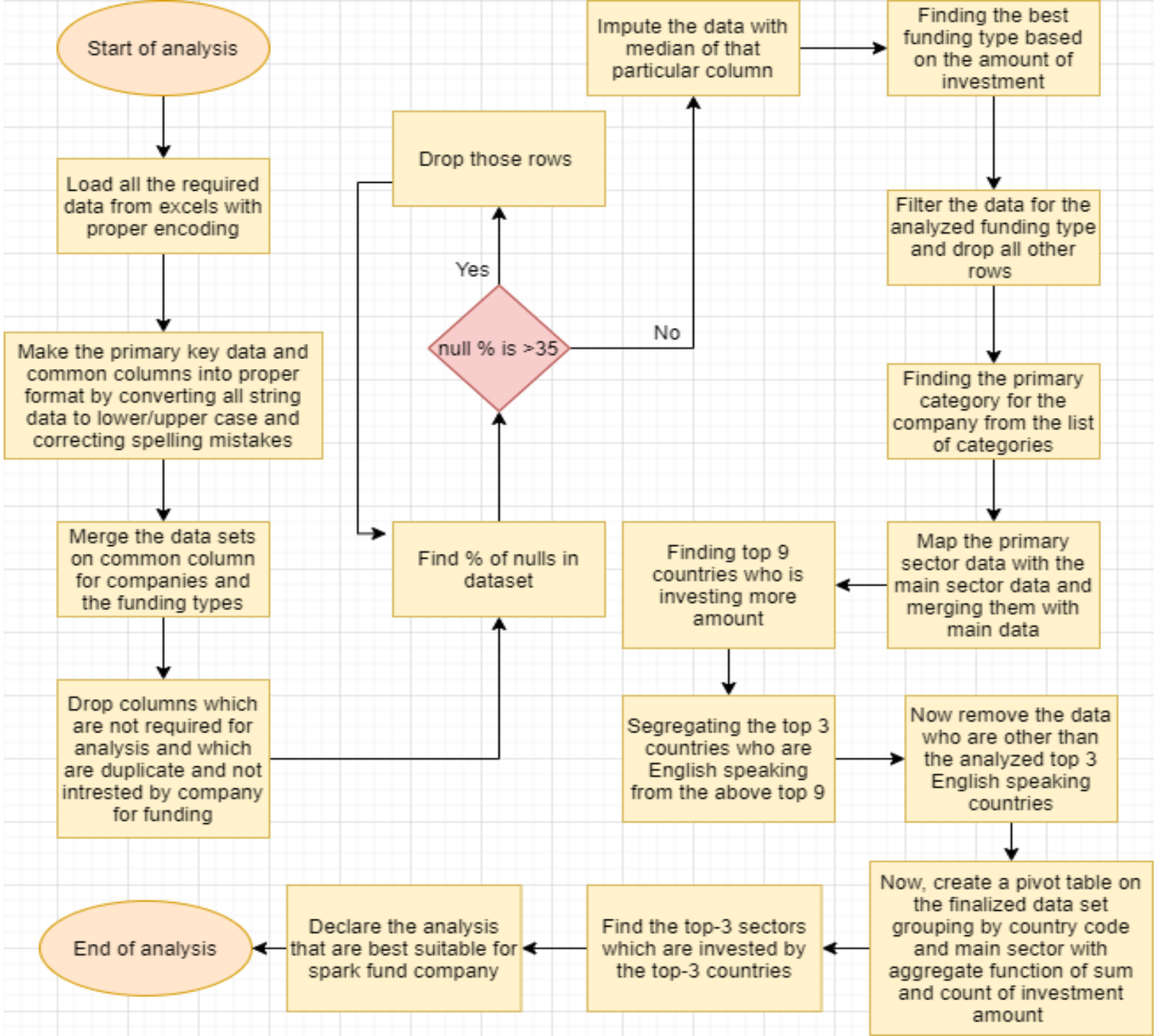
Name:

SRIKANTH PADMANABHUNI

Abstract :

- A company named '**Spark Fund'** is an asset management company, wants to make investment in few companies, before that Company wants to understand the global trends in investments so that it can take the investment decisions efficiently
- The objective is to identify the best sectors, countries and a sustainable investment type for making the investments
- The overall goal is to invest where others are investing and the best sectors, countries where most of the investors are investing.
- **Constrains of Spark Fund Company for investment:**
  - It wants to invest in between 5 to 15 million USD
  - It wants to invest only on English speaking countries for easy of communication.
- For finding the best countries and best sector for Spark Fund company, we are provided with three different datasets 'companies', 'mapping', 'rounds2'. Where
  - 'companies' dataset contains list of companies data with its country details.
  - 'mapping' dataset contains data about list of categories of company and its main sector to work with.
  - 'rounds2' dataset contains data about the funding type and the amount invested on that particular fund type.
- We need to find the best English speaking countries who invests on different funding types in different sectors with an investment amount in between 5-15 Million USD.
- We follow the following steps to analyse the data:
  1. Find the best investment amount required for different funding types so that Spark Fund can choose the best fitting for its investment amount.
  2. Finding the top 3 English speaking countries which invests on the above analysed funding type for budget range of Spark Fund investment.
  3. Finding the best sectors invested by the above analysed top three English speaking countries so that Spark Fund can invest on that particular sectors.

Problem Solving Methodology:



Analysis:

- The following are the analysis that are taken in different stages of the data analysis:

1. Data Cleaning:

- After loading the data, for primary key columns of data set ‘company_permalink’ and ‘permalink’ are having the error like two separate columns are different because two columns data are different in case of the letters which is not correct so corrected them by converting all of them to lower case letters
- After converting them to completely lower case, try to find the unique data and its count. During that, it was found some junk or Gibberish data in dataset so converted such data into normal English format by encoding it to the ‘ISO-8859-1’ and decoding it using ‘ASCII’.
- In ‘mapping’ dataset, I found that some data like ‘na’ characters are mis entered as ‘0’. So, replaced those ‘0’ with ‘na’

2. Merging Datasets:

- Merging ‘companies’ and ‘rounds2’ datasets:
  - Merging data on left join with common columns **‘company_permalink’, ‘permalink’** of ‘rounds2’, ‘companies’ datasets respectively.
  - After merging, lets drop the columns which are not required for our analysis like **'funding_round_code', 'funding_round_permalink', 'funded_at','permalink', 'homepage_url', 'state_code', 'region', 'city', 'founded_at','status’** .
  - After removing the un necessary columns it was found that required columns **‘raised_amount_usd’** column is having **’17.39’**, **‘category_list’** is having **‘2.97’, ‘country_code’** is having **‘7.55’** percentage of null data.
  - Among them ‘country_code’, ‘category_list’ are the categorical data which we need to impute by mode. But imputing with mode can give us the biased data and also the percentage of null data is negligible so lets drop those rows whose country_code or category_list data is NULL.
  - After dropping the above null data, percentage of null data for column **‘raised_amount_usd’** got reduced to **’13.92’**
  - Now, we can impute the data with either mean/median of data. But mean can cause the biased/outliers for the data. So, we can impute with median which will not occur any kind of outliers in data

- Before imputing that lets segregate the data further only with required four funding types **'venture', 'seed', 'angel', 'private_equity'** so that there might be the chance of reducing the percentage of null in **'raised_amount_usd'** column.
- After dropping the rows other than the funding types of 'venture/seed/angel.private_equity' the percentage of null reduced from **'13.92'** to **'9.4' .**
- Now, lets impute the null values of raised_amount_usd with the median such that, median of the particular funding type that column is pointing to. So, that the analysis will be more accurate than imputing with the median of whole data.

3. Finding the best funding_types:

- Lets create a data frame grouping by **'funding_types'** and the column data as **'raised_amount_usd'** with aggregate function of **'median'** (Why median? : Since, it acts as representative amount for the whole category).
- After grouping data, it was found that funding type **'venture'** is in the investment range(5-15 Million USD) of the company. So, it was concluded that funding type **'venture'** is best suitable for the company.
- So, drop the data which are other than the funding type **'venture'** and drop the column **'funding_type'** since it will have only one kind of data in it.

4. Finding the top three English speaking countries:

- First sorting the countries who made total highest investments and extracting the top nine among them.
- From top nine finding the top three English speaking countries among them.
- On analysing, it was found that countries with country code **'USA', 'GBR', 'IND'** are top three English speaking and highly invested companies in market
- Now, clean the data again such that drop the rows which are not in the top three English countries.

5. Finding the primary_sector from the category_list and mapping with 'mapping' dataset:

- As per the company guidelines from the list of categories first denoted category is primary and important than others. So, created a column with **primary_sector** from the category_list
- Before mapping, created a column with name **'main_sector'** in mapping.csv such that each category in the row gives us the 'sector' it was focused on using melt in pandas
- After that, mapped master data with the mapping data set and plotted the graph for different categories and its count

6. Segregate the data based on the lower and upper bounds of Investment amount:

- Lets reduce the data by dropping the rows whose investment amount is not in range of companies investment amount for analysing the data further easily

7. Creating final data frames for each country (D1, D2, D3):

- **D1:**
  - Created a dataset D1 which contains data of top1 English speaking country 'USA'.
  - Created a another dataset which contains data that is grouped by 'main_sector' and 'raised_amount_usd' with the aggreagtions of 'sum' and 'count'.
  - Merged about two data sets on 'main_sector' column.
  - On doing above we will get the data of each sector and the total amount invested and the total number of investments made on that sector.
- Followed the same fashion of strategy for the other top countries 'GBR' and 'IND'.
- By creating the above data sets, I analysed the total number of investments and the total sum of investments made by the top three countries.
- Later created the bar plots for the each data set of D1, D2, D3 as follows:
  - For Sector and the total number of investments made for that sector.
  - For Sector and the sum of investments made for that sector.
- On plotting above plots, I have analysed the top 3 sectors for each Data set for total investments and the total sum of investments.

- Finally for finding the top company which invested more in the top 2 sectors of each country:
  - Created a pivot table with index 'raised_amount_usd' and index as 'company_permalink' (Why company_permalink? Since it is the unique key to find the company uniquely)
  - By plotting graphs for above data, I have analysed the top company which received the highest/total number of investments.
- For the better and easy understanding, plotted a single graph for both total number of investments and the total sum of investments on each sector for each country.

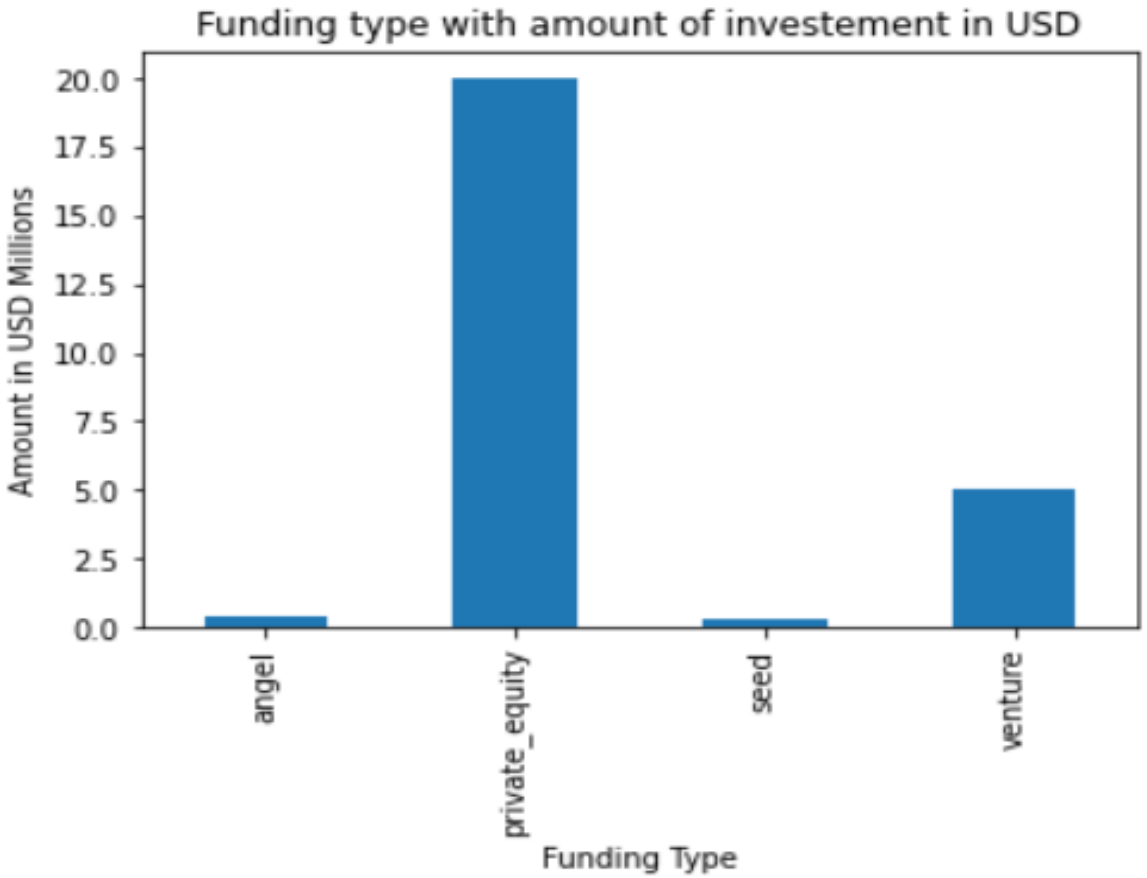
Results:

1. A plot showing **the representative amount of investment** in each funding type. This chart should make it clear that a certain funding type (FT) is best suited for Spark Funds:

- This graph clearly shows that funding_type:

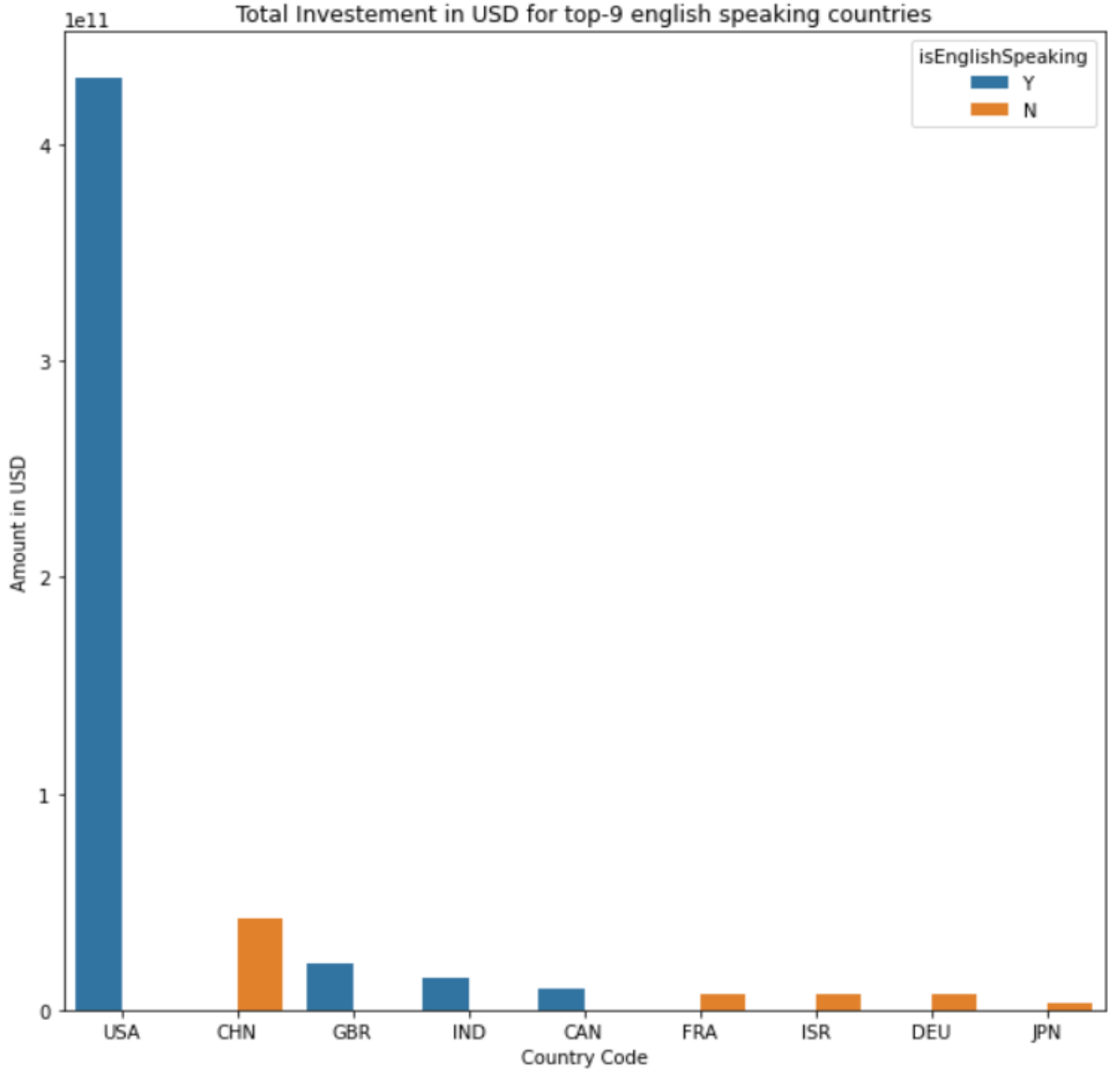
1. **private_equity** type requires the representative investment amount more than 15 Million USD which is out of range of the budget of Spark Fund
2. **angel** & **seed** type requires less representative investment amount than 5 Million USD dollars which is again less than the lower boundary of Spark Fund
3. **venture** type requires the representative investment amount of 5 Million USD which is in range of the budget of Spark Fund

* Hence, by this graph we can depict that **Venture** funding type is best suitable for Spark Fund comapny



2. A plot showing the top 9 countries against the total amount of investments of funding type FT. This should make the top 3 countries (Country 1, Country 2, and Country 3) very clear:

- This graph depicts the top 9 countries with differentiating English and Non-English speaking and the total sum of investments made by those countries.
- This graph clearly shows top 3 English speaking countries who invested more are **'USA', 'GBR', 'IND'.**



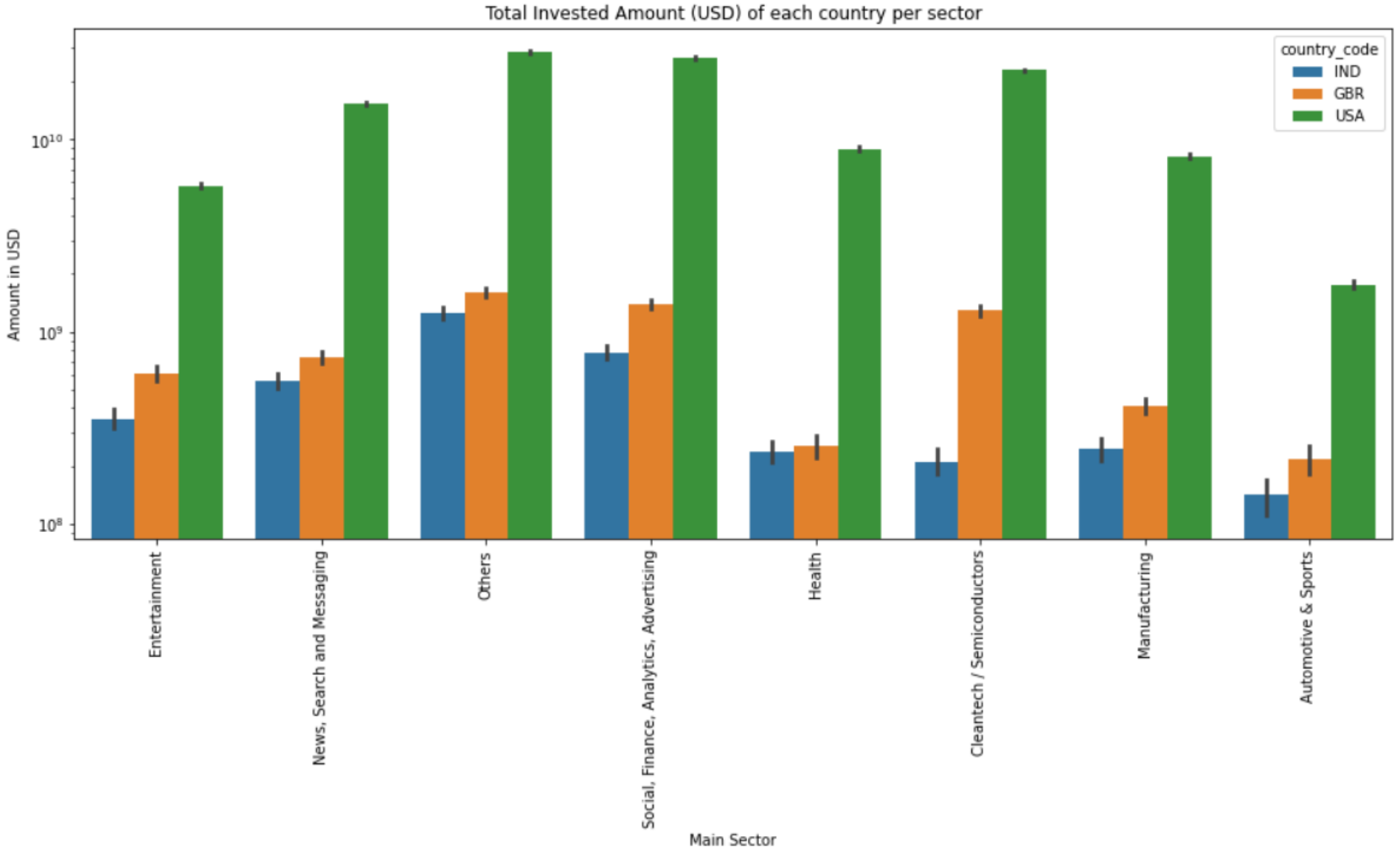
3. A plot showing the top 9 countries against the total amount of investments of funding type FT. This should make the top 3 countries (Country 1, Country 2, and Country 3) very clear:

- Here top 3 countries are

  'USA', 'GBR', 'IND'.

- Using log scale in y-axis for amount in USD for better understanding of bars.
- This graph clearly shows countries having the top3 sectors with highest investments
- And it is clearly visible that top3 sectors with highest investment foreach country are as follows:

1. USA & GBR:
   1. Others
   2. Social, Finance, Analytics, Advertising
   3. Cleantech/Semiconductors
2. IND:
   1. Others
   2. Social, Finance, Analytics, Advertising
   3. News, Search and Messaging



Conclusion:

- Based on the analysis made we can conclude that Spark Fund can make an investment of 5-15 Million USD for Funding type of Venture

top 3 English speaking countries(USA, GBR, IND) invested on the following top 3 sectors as follows:

- USA:

1. Others
2. Social, Finance, Analytics, Advertising
3. Cleantech/Semiconductors

- GBR:

1. Others
2. Social, Finance, Analytics, Advertising
3. Cleantech/Semiconductors

- IND:

1. Others
2. Social, Finance, Analytics, Advertising
3. News, Search and Messaging

- Therefore, **Spark Fund can invest on** top two sectors **'Others'** and **'Social, Finance, Analytics, Advertising'**.
- If 'Others' is not considered as sector, then it can invest on **'Cleantech/Semiconductors'** since it is the Third top sector of USA and GBR countries.