AMITY UNIVERSITY
UTTAR PRADESH

Assignment No 1

Sub Name: Fundamental of Data Analytics **Faculty Name** : Dr K.K. Singh

Submission Date:25th August 2022 **Maximum Marks**:20

Problem Statement

Objectives

Note: This is an individual assignment.

Project Brief

You work for Spark Funds, an asset management company. Spark Funds wants to make investments in a few companies. The CEO of Spark Funds wants to understand the global trends in investments so that she can take the investment decisions effectively.

Business and Data Understanding

Spark Funds has two minor constraints for investments:

1. It wants to invest between **5 to 15 million USD** per round of investment
2. It wants to invest only in **English-speaking countries** because of the ease of communication with the companies it would invest in

- For your analysis, consider a country to be English speaking only if English is one of the official languages in that country
- You may use this list: Click here for a list of countries where English is an official language.

These conditions will give you sufficient information for your initial analysis. Before getting to specific questions, let's understand the problem and the data first.

1. What is the strategy?

Spark Funds wants to invest where most **other investors are investing**. This pattern is often observed among early stage startup investors.

2. Where did we get the data from?

We have taken real investment data from **crunchbase.com**, so the insights you get may be incredibly useful. For this assignment, we have divided the data into the following files:

You have to use three main data tables for the entire analysis (**available for download on the next page**):

3. What is Spark Funds' business objective?

The business objectives and goals of data analysis are pretty straightforward.

1. **Business objective:** The objective is to identify the best sectors, countries, and a suitable investment type for making investments. The overall strategy is to invest where others are investing, implying that the 'best' sectors and countries are the ones 'where most investors are investing'.
2. **Goals of data analysis**: Your goals are divided into three sub-goals:
    1. **Investment type analysis**: Comparing the typical investment amounts in the venture, seed, angel, private equity etc. so that Spark Funds can choose the type that is best suited for their strategy.
    2. **Country analysis**: Identifying the countries which have been the most heavily invested in the past. These will be Spark Funds' favourites as well.
    3. **Sector analysis**: Understanding the distribution of investments across the eight main sectors. (Note that we are interested in the eight 'main sectors' provided in

the **mapping file**. The two files — **companies and rounds2** — have numerous sub-sector names; hence, you will need to map each sub-sector to its main sector.)

4. How do you approach the assignment? What are the deliverables?

The entire assignment is divided into checkpoints to help you navigate. For each checkpoint, you are advised to fill in the tables into the spreadsheet provided in the download segment. The tables are also mentioned under the **'Results Expected'** section after each checkpoint. Since this is the first assignment, you have been provided with some additional guidance. Going forward you will be expected to structure and solve the problem by yourself, just like you would be solving problems in real life scenarios.

Important Note: All your code has to be submitted in one Jupyter notebook. For every checkpoint, keep writing code in one well-commented Jupyter notebook which you can submit at the end.

Downloads

1. Company details

companies: A table with basic data of companies

Description of Companies Table	
Attribute	Description
Permalink	Unique ID of company
name	Company name
homepage_url	Website URL

Description of Companies Table	
Attribute	Description
category_list	Category/categories to which a company belongs
status	Operational status
country_code	Country Code
state_code	State

You can download the companies data here.

Companies Download

2. Funding round details:

rounds2: The most important parameters are explained below:

Description of rounds2 Table	
Attributes	Description
company_permalink	Unique ID of company
funding_round_permalink	Unique ID of funding round
funding_round_type	Type of funding – venture, angel, private equity etc.
funding_round_code	Round of venture funding (round A, B etc.)
funded_at	Date of funding
raised_amount_usd	Money raised in funding (USD)

Rounds2 Download

3. Sector Classification:

mapping.csv: This file maps the numerous **category names** in the companies table (such 3D printing, aerospace, agriculture, etc.) to eight broad **sector names.** The purpose is to simplify the analysis into eight sector buckets, rather than trying to analyse hundreds of them.

Mapping Download

4. **Excel File (mandatory submission):** Download the Excel spreadsheet from below. It contains all the tables you need to fill in.

Investments Download

5. **Presentation template (mandatory submission):** Download the sample PPT from below. The structure is a suggestion; make sure not to exceed 10 slides. Once your presentation is ready, convert the document in PDF format for submission.

Spark Funds Presentation Download

Checkpoints - Part 1

Checkpoint 1: Data Cleaning 1

1. Load the companies and rounds data (provided on the previous page) into two data frames and name them **companies** and **rounds2** respectively.
2. Table 1.1: The table below is just for reference — you need to fill out the spreadsheet which is attached in the download section. This holds true for all the tables.

Results Expected: Table 1.1

Table 1.1: Understand the Data Set

How many **unique companies** are present in **rounds2**?	
How many **unique companies** are present in **companies**?	
In the **companies** data frame, which column can be used as the unique key for each company? Write the **name of the column**.	
Are there any companies in the rounds2 file which are not present in companies? Answer yes or no: **Y/N**	

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?	

After this, you will need to work only with the **master frame.**

Checkpoint 2: Funding Type Analysis

This is the first of the three goals of data analysis – investment type analysis.

The funding types such as seed, venture, angel, etc. depend on the type of the company (startup, corporate, etc.), its stage (early stage startup, funded startup, etc.), the amount of funding (a few million USD to a billion USD), and so on. For example, seed, angel and venture are three common stages of startup funding.

- Seed/angel funding refer to early stage startups whereas venture funding occurs after seed or angel stage/s and involves a relatively higher amount of investment.
- Private equity type investments are associated with much larger companies and involve much higher investments than venture type. Startups which have grown in scale may also receive private equity funding. This means that if a company has reached the venture stage, it would have already passed through the angel or seed stage/s.

Spark Funds wants to choose one of these four investment types for each potential investment they will make.

Considering the constraints of Spark Funds, you have to decide one funding type which is most suitable for them.

1. Calculate the **most representative value of the investment amount** for each of the four funding types (venture, angel, seed, and private equity) and report the answers in **Table 2.1**
2. Based on the most representative investment amount calculated above, which investment type do you think is the most suitable for Spark Funds?

Considering that Spark Funds wants to invest between **5 to 15 million USD** per investment round, which investment type is the most suitable for it? Identify the investment type and, for further analysis, filter the data so it only contains the chosen investment type.

Checkpoints - Part 2

Checkpoint 3: Country Analysis

This is the second goal of analysis — **country analysis**.

Now that you know the type of investment suited for Spark Funds, let's narrow down the countries.

Spark Funds wants to invest in countries with the highest amount of funding for the chosen investment type. This is a part of its broader strategy to invest where **most investments are occurring**.

1. Spark Funds wants to see the top nine countries which have received the highest total funding (across ALL sectors for the chosen investment type)
2. For the chosen investment type, make a data frame named **top9** with the top nine countries (based on the total investment amount each country has received)

Identify the top three English-speaking countries in the data frame top9.

Results Expected: All codes for data frame top9. Fill out Table 3.1.

Table 3.1: Analysing the Top 3 English-Speaking Countries

1. Top English-speaking country	
2. Second English-speaking country	
3. Third English-speaking country	

Now you also know the three most investment-friendly countries and the most suited funding type for Spark Funds. Let us now focus on finding the best sectors in these countries.

Checkpoint 4: Sector Analysis 1

This is the third goal of analysis — **sector analysis.**

When we say sector analysis, we refer to one of the **eight main sectors** (named **main_sector**) listed in the mapping file (note that 'Other' is one of the eight main sectors). This is to simplify the analysis by grouping the numerous category lists (named 'category_list') in the mapping file. For example, in the mapping file, category_lists such as '3D', '3D Printing', '3D Technology', etc. are mapped to the main sector 'Manufacturing'.

Also, for some companies, the category list is a list of multiple sub-sectors separated by a pipe (vertical bar |). For example, one of the companies' category_list is Application Platforms|Real Time|Social Network Media.

You discuss with the CEO and come up with the **business rule** that the first string before the vertical bar will be considered the **primary sector**. In the example above, ‘Application Platforms’ will be considered the primary sector.

1. **Extract** the primary sector of each category list from the **category_list column**
2. Use the **mapping file** 'mapping.csv' to map each primary sector to one of the eight main sectors (Note that ‘Others’ is also considered one of the main sectors)

Expected Results: Code for a merged data frame with each primary sector mapped to its main sector (the primary sector should be present in a separate column).

Checkpoint 5: Sector Analysis 2

Now you have a data frame with each company’s main sector (main_sector) mapped to it. When we say sector analysis, we refer to one of the eight main sectors.

Also, you know the top three English speaking countries and the most suitable funding type for Spark Funds. Let’s call the three countries 'Country 1', 'Country 2' and 'Country 3' and the funding type 'FT'.

Also, the range of funding preferred by Spark Funds is **5 to 15 million USD**.

Now, the aim is to find out the most heavily invested main sectors in each of the three countries (for funding type FT and investments range of 5-15 M USD).

1. Create three separate data frames D1, D2 and D3 for each of the three countries containing the observations of funding type FT falling within the 5-15 million USD range. The three data frames should contain:
   - All the columns of the master_frame along with the primary sector and the main sector
   - The total number (or count) of investments for each main sector in a separate column
   - The total amount invested in each main sector in a separate column

Using the three data frames, you can calculate the total number and amount of investments in each main sector.

Result Expected

1. Three data frames **D1, D2** and **D3**
2. Table 5.1: Based on the analysis of the sectors, which main sectors and countries would you recommend Spark Funds to invest in? Present your conclusions in the presentation. The conclusions are subjective (i.e. there may be no 'one right answer'), but it should be based on the basic strategy — invest in sectors where most investments are occurring.

Note: In the following table, all the observations refer to investments of the type FT within 5-15 M USD range.

Table 5.1 : Sector-wise Investment Analysis

	Country 1	**Country 2**	**Country 3**
1. Total number of investments (count)			
2. Total amount of investment (USD)			
3. Top sector (based on count of investments)			

	Country 1	Country 2	Country 3
4. Second-best sector (based on count of investments)			
5. Third-best sector (based on count of investments)			
6. Number of investments in the top sector (refer to point 3)			
7. Number of investments in the second-best sector (refer to point 4)			
8. Number of investments in the third-best sector (refer to point 5)			
9. For the top sector count-wise (point 3), which company received the highest investment?			
10. For the second-best sector count-wise (point 4), which company received the highest investment?			

Checkpoint 6: Plots

As a final step, you have to present your findings to the CEO of Spark Funds. Specifically, she wants to see the following plots:

1. A plot showing the fraction of total investments (globally) in venture, seed, and private equity, and the average 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.

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.
3. A plot showing the number of investments in the **top 3 sectors** of the **top 3 countries** on one chart (for the chosen investment type FT).

This plot should clearly display the top 3 sectors each in Country 1, Country 2, and Country 3.

Expected Result: The three plots.

Evaluation Rubric

Evaluation Rubric		
Criteria	Meets expectations	Does not meet expectations
Data understanding and preparation (10%)	All data quality issues are correctly identified and reported. The unique keys and number of unique entries are correctly identified. The files are collated correctly to create a master file.	Data quality issues are overlooked or are not identified correctly. Unique keys or values are not understood/identified correctly. The master file is not created / incorrectly created.

Evaluation Rubric		
Criteria	Meets expectations	Does not meet expectations
Cleaning and manipulating data (25%)	Data quality issues are addressed in the right way (missing value treatment etc.). If applicable, data is converted to a suitable and convenient format to work with using the right methods. Manipulation of dates and strings, if required, is done using correct and concise techniques/code.	Data quality issues are not addressed correctly. The variables are not converted to an appropriate format for analysis. The format of data is not altered to a convenient one and as a result, the analysis is done using longer methods / involves complex steps. String and date manipulation is not done correctly or is done using complex methods.
Data analysis (35%)	The analysis has a clear structure and the flow is easy to understand. The funding, country and sector wise analysis are done correctly and according to the instructions. Appropriate realistic assumptions are made wherever required. The use cases of aggregation, drill down, slicing, dicing	The analysis lacks a clear structure and is not easy to follow. The three types of analysis are not conducted correctly and the results are incorrect. Realistic assumptions are not made wherever required or unrealistic ones are made.

Evaluation Rubric		
Criteria	Meets expectations	Does not meet expectations
	etc. operations are correctly identified and conducted in Python. The investment type, list of countries and the sectors is correct. Appropriate plots are created to present the results of the analysis. The choice of plots for respective cases is correct. The plots should clearly present the relevant insights and should be easy to read. The axes and important data points are labelled correctly.	The aggregation, drill down, slicing, dicing etc. operations are not performed correctly. The investment type, list of countries and the sectors is incorrect. All relevant plots are not created. The choice of plots is not ideal and the plots are either difficult to interpret or lack clarity or neatness. Relevant insights are not clearly presented by the plots. The axes and important data points are not labelled correctly / are not neatly labelled.
Presentation of results (20%)	The presentation has a clear structure, is not too long and explains the most important results concisely.	The presentation lacks structure, is too long or does not put emphasis on the important observations. Contains unnecessary details or lacks the important ones.

Evaluation Rubric		
Criteria	Meets expectations	Does not meet expectations
	If any assumptions are made, they are stated clearly.	Assumptions made, if any, are not stated clearly.
Conciseness and readability of the code (10%)	The code is concise and syntactically correct. Wherever appropriate, built-in functions are used instead of writing long code (if-else statements, for loops). The code is readable with variables appropriately named and detailed comments are written wherever necessary.	Long and complex code used instead of shorter built-in functions. Code readability is poor because of vaguely named variables or lack of comments wherever necessary.