**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) |  |  |  |
| 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. |
| 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 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 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.    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.    If any assumptions are made, they are stated clearly. | The presentation lacks structure, is too long or does not put emphasis on the important observations.    Contains unnecessary details or lacks the important ones.    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. |  |

# Gradeded Old Questions

### Seating Arrangement

**Description**

N people sit around a circular table. You have to find the probability that two particular people won't be sitting together.  
  
The input will have the number N and the output should have the probability printed as a float type number rounded off to four decimal places.  
  
**Hint:** This problem follows the logic of circular permutation which you can read about [here.](https://www.dyclassroom.com/aptitude/circular-permutation)  
  
Please note that if the answer is 3.6, it is to be reported as 3.6 and not 3.6000  
  
Sample Input:  
15  
  
Sample Output:

0.8571

### Coloured cards

**Description**

A card stack contains white and black cards. Two cards are drawn randomly without replacement. The probability of selecting a white and then a black card is x. The probability of selecting a white card in the first draw is y. You have to find the probability of drawing a black card, given that the first card drawn was white.  
The input will contain two lines with x and y respectively.    
The output should be displayed as a float(no need to round it off).  
Sample Input:  
0.2  
0.5  
Sample Output:  
0.4

### Entrance Test

**Description**

Consider that the scores of an entrance test are normally distributed.   
If the test has a mean of m and a standard deviation of s.  
You have set the cut off as 90 marks for the test. What is the probability of a student passing the test?  
The input will contain m and s in two lines in the same order. The output has to be rounded off to two decimal places.  
  
Please note that if the answer is 3.6, it is to be reported as 3.6 and not 3.60  
  
Sample Input:  
100  
10  
Sample Output:  
0.84