**OBJECTIVES**

**Project Brief:**You are working for ‘X’ Funds, an asset management company. ‘X’ Funds wants to make investments in a few companies. The CEO of ‘X’ Funds wants to understand the global trends in investments so that he/she can take the investment decisions effectively.  
  
**Business and Data Understanding:**‘X’ Funds have two minor constraints for investments:  
1. They want to invest between **5 to 15 million USD** per round of investment.  
2. They want to invest only in **English-speaking countries** because of the ease of communication with the companies they’d invest in.  
3. 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**](http://www.emmir.org/fileadmin/user_upload/admission/Countries_where_English_is_an_official_language.pdf) to know the name 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 data first.

| **1. What is the Strategy?** The ‘X’ Funds Company wants to invest where most **other investors are investing**. This pattern is often observed among early stage start-up 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.  You have to use 3 data files for the entire analysis: ***1. companies.txt***: A text file with basic data of companies. **1. Attributes description of companies.txt file** | | | |
| --- | --- | --- | --- |
| Attributes | Description | | |
| Permalink | Unique ID of company | | |
| name | Company name | | |
| homepage\_url | Website URL | | |
| category\_list | Category/categories to which a company belongs | | |
| status | Operational status | | |
| country\_code | Country | | |
| state\_code | State | | |
| ***2. rounds2.csv***: A csv file with data about investments. The most important parameters are explained below: **2. Attribute description of rounds2.csv file** | | |
| 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.) |

***3. mapping\_file.csv***: This file maps the numerous sector names (like 3D printing, aerospace, agriculture etc.) to 8 main sector names. The purpose of having 8 main sectors is to simplify the analysis into 8 sector buckets, rather than trying to analyze hundreds of them.   
**3. What is ‘X’ 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 investments are happening.
2. **Goals of data analysis**: Your goals are divided into 3 main sub-goals:
   * **Investment type analysis**: Understanding investments in venture, seed/angel, private equity categories etc. so ‘X’ Funds can decide which type is best suited for their strategy.
   * **Country analysis**: Understanding which countries have had the most investments in the past. These will be ‘X’ Funds’ favorites as well.
   * **Sector analysis**: Understanding the distribution of investments across the 8 main sectors (note that we are interested in the 8 main sectors provided in the mapping file. The 2 files, companies and rounds2, have numerous sub-sector names; hence you will need to map each sub-sector to its main sector).

**DATA CLEANING 1**

1. Load the two files **companies.txt and rounds2.csv** into two data frames and name them ***companies***and***rounds2*** respectively.
2. Table 1.1: The table below is just for reference, Fill out the values as you get.

**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 **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: DATA CLEANING**

1. Replace all the NA values from the ***raised\_amount\_usd***column of the ***master frame***.

**Table 2.1:**NA Values Treatment

|  |  |
| --- | --- |
| How many NA values are present in the column ***raised\_amount\_usd***? |  |
| What do you replace NA values of ***raised\_amount\_usd*** with? Enter a numeric value. |  |

**FUNDING TYPE ANALYSIS**This is the first of the three goals of data analysis – investment-type analysis.

The funding types like seed, venture, angel etc. depend upon the type of the company (start-up, corporate etc.), its stage (early stage start-up, funded start-up etc.), the amount of funding (a few million USD to a billion USD) etc. For example, seed, angel and venture are three common stages of start-up funding.

* Seed / angel funding refer to early stage start-ups whereas venture funding occurs after seed / 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. Start-ups which have grown in scale may also receive private equity funding. This means if a company has reached venture stage, it would have already passed through angel / seed stage(s).

**‘X’ Funds wants to choose one of these four investment types for each potential investment they will make.**Considering the constraints of ‘X’ Funds, you have to decide one funding type most suitable for them.

1. Calculate the average investment amount for each of the four funding types (venture, angel, seed and private equity) and report the answers in **Table 3.1.**
2. Based on the average investment amount calculated above, which investment type do you think is the most suitable for ‘X’ Funds?

**Table 3.1**: Average values of investments for each of these funding types

|  |  |
| --- | --- |
| Average funding amount of **venture type** |  |
| Average funding amount of **angel type** |  |
| Average funding amount of **seed type** |  |
| Average funding amount of **private equity type** |  |
| Considering that ‘X’ Funds wants to invest between **5 to 15 million USD** per investment round, which investment type is the most suitable for them? |  |

**COUNTRY ANALYSIS**

This is the second goal of analysis – **country  analysis**.  
Now that you know the type of investment suited for ‘X’ Funds, let us narrow down the countries.   
‘X’ Funds wants to invest in countries having the highest amount of funding for the chosen investment type. This is a part of their broader strategy to invest where **most investments are occurring**.

1. ‘X’ Funds wants to see the top 9 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 top9 countries (based on the total investment amount each country has received).

**Identify the top 3 English speaking countries in the data frame top9.**

Table 4.1:Analyzing top 3 English speaking countries

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

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

**SECTOR ANALYSIS1**This is the third goal of analysis**– sector analysis.**When we say sector analysis, we refer to one of the **8 main sectors**(named ***main\_sector***) listed in the mapping file (note that ‘Other’ is one of the 8 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 like ‘3D’, ‘3D Printing’, ‘3D Technology’ etc. are mapped to the main sector ‘Manufacturing’.  
Also, for some companies, 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.  
The **business rule** tells you 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 **mapping file** "*mapping.csv"* to map each primary sector to one of the 8 main sectors. (Note that ‘Others’ is also considered one of the main sector)  
**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). Also,

**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 8 main sectors.  
Also, you know the top 3 English speaking countries and the most suitable funding type for ‘X’ 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 ‘X’ 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 range of investment between 5-15 M USD).

1. Create three separate data frames D1, D2 and D3 for each of the 3 countries containing the observations of funding type FT falling between 5 to 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.   
**Note: In the following table, all observations refer to investments of the type FT between 5 to 15 M USD.**  
Table 6.1: Sector-wise Investment Analysis

|  | **Country1** | **Country2** | **Country3** |
| --- | --- | --- | --- |
| 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. Third sector name (no. of investment-wise) |  |  |  |
| 6. Number of investments in top sector (3) |  |  |  |
| 7. Number of investments in second sector (4) |  |  |  |
| 8. Number of investments in third sector (5) |  |  |  |
| 9. For point 3 (top sector count-wise), which company received the highest investment? |  |  |  |
| 10. For point 4 (second best sector count-wise), which company received the highest investment? |  |  |  |

**PLOTS**As a final step, you have to present your findings to the CEO of ‘X’ Funds.

1. **One pie chart** 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 ‘X’ Funds.
2. **One bar chart** showing 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. **Any chart type** you think is suitable: This should show 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.