CASE STUDY 2

CODING NINJAS PROJECT

OCTOBER 10, 2022

HARSH MALIK

QUESTION 1.

1. In this question, I have first dropped the NaN values (if any) from the 'CityLocation' column in the Data Frame since they are of no use anyway.
2. Next, I have changed some of the values to the default notation specified in the question, such as:
   - Changed "Delhi" to "New Delhi"
   - Changed "bangalore" to "Bangalore"
   - Checked if any other of the selected cities need to be processed so that there are no errors when the answer is being calculated
3. Now, I have created a list with the names of the cities whose frequency we need to calculate. This is done to filter and speed up the process when both of these (list of names of cities and 'CityLocation' Series) are passed to the function I have written which will help in calculating the frequency of each location.
4. The function created for this question is a simple dictionary creator function where we iterate over each row of the 'CityLocation' column and then examine its data as follows:
   - Firstly, we check whether there is a single entry or multiple entries.
   - If it is a single entry, we then simply check if it is one of the cities that we are looking for. If so, we increase its frequency in the dictionary, else we move on.
   - If there are multiple entries, we first use the .split('/') function to split the multiple values separated by '/' in the row. Then we apply the .strip() function to remove any spaces before and after the value that we have received after extraction using the .split() function.
   - After this is done, again we check if it is one of the locations that we are looking for and increase its frequency or ignore accordingly.
5. Now after the processing is done, we simply plot the bar graph as asked in the question using the dictionary keys and values that our function has evaluated and get our final answer, i.e., which city is the most promising one for the startup of our friend.

QUESTION 2.

1. Data cleansing is done first; undisclosed investors are eliminated from the data frame.
2. Now, I have created a function that returns a dictionary with consisting of names of investors as keys and their number of investments as values.
3. Now this dictionary is sorted in descending order with respect to the values present in it.
4. Now, we have created 2 empty lists; l1 and l2 and run a loop 5 times over the sorted dictionary in order to iterate and retrieve the top values in the dictionary.
    a. l1 consists of the names of the investors.
    b. l2 consists of the values of their number of investments.
5. The values are printed while iterating the dictionary.
6. To show the difference among the investors more effectively and in a more visual manner, a pie chart has been plotted which shows the different percentages of the number of investments made by the respective investors.

QUESTION 3.

1. In this question as well, first data cleaning is performed; spelling corrections, removal of undisclosed investors and NaN values from the concerned columns in the data frame.
2. This time I have created a new function, where whenever an investor has invested in the same start-up again in a different round, it has not been counted as a new investment and therefore it is ignored. A dictionary consisting of names of various investors as keys and a list of different start-ups they have invested in is their value (where no value is repeated) is created using this function.
3. Now we simply calculate the length of the list for each key and get the top 5 key value pairs. Hence, our final list of top investors has been evaluated.
4. A bar chart is now plotted for this data to show the difference among the top 5 investors visually. This helps us illustrate the difference in numbers among the top investors.

QUESTION 4.

1. Same as other questions, first data cleansing is performed; spell checks, removal of undisclosed investors and NaN values from the concerned columns in the data frame.
2. Now, I have extracted only those rows where the Investment Type is either Seed Funding or Crowd Funding, as these are the main investment types that will help our friend's start-up.
3. Now the same process of separating wheat from chaff is done as shown in question 3, using the same functions and process again.
4. A bar chart is plotted at the end to illustrate the differences in the data retrieved after processing the data frame.

QUESTION 5.

1. This question is exactly same as question 4 and same methods are followed. The only difference being that after data cleansing, only those rows where Investment Type is Private Equity are extracted from the data frame and rest are eliminated. Everything else remains the same.
2. A bar chart is plotted at the end.