Explanations

Question 1

Your Friend has developed the Product and he wants to establish the product startup and he is searching for a perfect location where getting the investment has a high chance. But due to its financial restriction, he can choose only between three locations - Bangalore, Mumbai, and NCR. As a friend, you want to help your friend deciding the location. NCR include Gurgaon, Noida and New Delhi. Find the location where the most number of funding is done. That means, find the location where startups has received funding maximum number of times. Plot the bar graph between location and number of funding. Take city name "Delhi" as "New Delhi". Check the case-sensitiveness of cities also. That means, at some place instead of "Bangalore", "bangalore" is given. Take city name as "Bangalore". For few startups multiple locations are given, one Indian and one Foreign. Consider the startup if any one of the city lies in given locations.

Explanation:

Here "loc" variable holds all rows of "CityLocation" column. I order to get all cities name where more than 1 city name in present "split" function is used. It split string present is each row on the basis of forward slash "/".
"a=loc.str.split("/",expand=True)" here variable "a" holds a dataframe with 2 columns. Column 1 holds the value to city name which comes first in a particular row and column 2 holds the values of second city. If 2nd city is not present then column 2 hold "None" values. Like this:

	0	1
0	Bangalore	None
1	Mumbai	None
2	New Delhi	None
3	Mumbai	None
4	Hyderabad	None
5	Bangalore	None
6	Ahmedabad	None
7	Gurgaon	None
8	Bangalore	None
9	Noida	None
10	Mumbai	None

Then in line 12 both the columns are concatenated to form a single column and in the same line extra spaces of each string in each row is removed and rows with "NA" values are dropped. This is stored in the variable "loc". Now this data is ready to get cleaned.

For Data cleaning and correcting the required cities name "Boolean array" is used. Here all city names are not considered, only cities present in Question are renamed and per requirement.

Then again "Boolean array" is used for converting "New Delhi", "Gurgaon" & "Noida" into "NCR". And also for removing other irrelevant cities from "loc" variable.

In line 27 value_counts() function is used for getting the count of unique cities present. And this is stored in "total_count" variable with is of type pandas series. Then "city" & "city_count" variables

are used for simply storing the index (key) and value of "total_count" and printing city names with its count value.

Bar graph is best suited for showing city name with their count values. On X-axis name of city is present and on Y-axis its count is present.

Conclusion:

From result it can be concluded that NCR will be the best location to establish the product startup. Because this is the location where the most number of funding is done and startups established here are likely to get funding easily. Then comes "Bangalore" then "Mumbai".

Question 2

Even after trying for so many times, your friend's startup could not find the investment. So you decided to take this matter in your hand and try to find the list of investors who probably can invest in your friend's startup. Your list will increase the chance of your friend startup getting some initial investment by contacting these investors. Find the top 5 investors who have invested maximum number of times (consider repeat investments in one company also). In a startup, multiple investors might have invested. So consider each investor for that startup. Ignore undisclosed investors.

Explanation:

Here "name" variable holds the values of "InvestorsName" column which is of type Pandas Series. Then all "NA" values of dropped from this.

"split" Function is used in order to extract all the names which are separated by commas. And this pandas series is converted into numpy array. Which is further flattened into 1-dimension with this piece of code: "(np.concatenate(all_name).flat)" and extra spaces in each string are removed with split function. Then finally this numpy array is converted into pandas series using pd.series function (line 15).

In line 18 value_counts() function is used for getting the count of unique names of investors. This is stored in "name_count" variable.Then simply printed top 5 investor's names with number of investments they have done.

Again Bar graph is best suited for showing Investor's name with their count values of number of investments. On X-axis name of investor is present and on Y-axis their count is plotted (Only top 5 investors).

Conclusion:

From result it is clear that "Sequoia Capital" has invested most number of times. So this investor will be the best for contacting for getting funds and investments. Then comes "Accel Partners", "Kalaari Capital", "SAIF Partners" and "Indian Angel Network".

Question 3

After re-analyzing the dataset you found out that some investors have invested in the same startup at different number of funding rounds. So before finalizing the previous list, you want to improvise it by finding the top 5 investors who have invested in different number of startups. This list will be more helpful than your previous list in finding the investment for your friend startup. Find the top 5 investors who have invested maximum number of times in different companies. That means, if one investor has invested multiple times in one startup, count one for that company. There are many errors in startup names. Ignore correcting all, just handle the important ones - Ola, Flipkart, Oyo and Paytm.

Explanation:

Firstly all columns except "InvestorsName" and "StartupName" have been dropped using "drop" function of pandas dataframe (line 8). Then "NA" values of column "InvestorsName" and "StartupName" have been drop and finally index are reset using "reset_index" function. Now dataset is ready to get cleaned.

Cleaned the important startup names like: Ola, Flipkart, Oyo and Paytm using boolean array.

"name" variable is used for "InvestorsName" and "split" Function is used in order to extract all the names which are separated by commas (line 28).

Two dictionaries named "d1" and "d2" are created.

Dict "d1" is for maintaining "investor name" and associated startup name: After removing extra leading space, for each investor names a set is maintained. That means for dict "d1" key is "Investor's name" and values are "startup name" which is inside set. Here sets are used purposely because sets cannot hold duplicate values. So if a investor has already invested in any startup then for the second time that startup name will not be added into set because it has already been added.

Dict "d2" is for maintaining "investor name" as key and length of set which contain startup name from dict "d1" as values. In line 40: "setdefault function" is use to insert a new key (ie. investor name) and values in the dictionary "d1" and if that key (investor name) is already present then it will update the value of that key without losing previous values. And "add" function is use to insert items in set.

Then finally dictionary d2 is sorted with respect to values using itemgetter function present in operator library. And top 5 items of this dictionary is printed. That means top 5 investor's name who have invested maximum number of times in different companies with their no of investment have been printed and plotted on the graph. X-axis: "investor_name" and Y-axis: "number of investment" of that investor.

Conclusion:

From result it is clear that "Sequoia Capital" has invested most number of times in different companies. So this investor will be the best for contacting for getting funds and investments. Then comes "Accel Partners", "Kalaari Capital", "Indian Angel Network" and "Blume Ventures".

Question 4

Even after putting so much effort in finding the probable investors, it didn't turn out to be helpful for your friend. So you went to your investor friend to understand the situation better and your investor friend explained to you about the different Investment Types and their features. This new information will be helpful in finding the right investor. Since your friend startup is at an early stage startup, the best-suited investment type would be - Seed Funding and Crowd funding. Find the top 5 investors who have invested in a different number of startups and their investment type is Crowd funding or Seed Funding. Correct spelling of investment types are - "Private Equity", "Seed Funding", "Debt Funding", and "Crowd Funding". Keep an eye for any spelling mistake. You can find this by printing unique values from this column. There are many errors in startup names. Ignore correcting all, just handle the important ones - Ola, Flipkart, Oyo and Paytm.

Explanation:

Firstly all columns except "InvestorsName", "InvestmentType" and "StartupName" have been dropped using "drop" function of pandas dataframe (line 8). Then "NA" values of column "InvestorsName", "InvestmentType" and "StartupName" have been drop and finally index are reset using "reset_index" function. Now dataset is ready to get cleaned.

First "InvestmentType" is cleaned and spelling errors have been removed using Boolean array. Then rows which are not required have been dropped. We only need "Seed funding" and "Crowd funding" for an early stage startup so "Private Equity" and "Debt funding" rows have been droped (line 25) and index are reset. Now StartupName column should be cleaned, only the important startup names like: Ola, Flipkart, Oyo and Paytm has been cleaned using boolean array.

"startup" variable is used for "StartupName" and "name" variable is used for "InvestorsName" and "split" Function is used in order to extract all the names which are separated by commas (line 40,41,42).

Rest other steps are same as in question 3 (From line 44 onwards).

Finally top 5 investor's name who have invested maximum number of times in different companies having "InvestmenType" as "Seed funding" and "Crowd funding" with their no of investment have been printed and plotted on the graph. X-axis: "investor_name" and Y-axis: "number of investment" of that investor.

Conclusion:

From result it is clear that "Indian Angel Network" has invested most number of times in different companies having "InvestmenType" as "Seed funding" and "Crowd funding". So this investor will be the best for contacting for getting funds and investments for an early stage startup. Then comes "Rajan Anandan", "LetsVenture", "Anupam Mittal" and "Kunal Shah".

Question 5

Due to your immense help, your friend startup successfully got seed funding and it is on the operational mode. Now your friend wants to expand his startup and he is looking for new investors for his startup. Now you again come as a saviour to help your friend and want to create a list of probable new new investors. Before moving forward you remember your investor friend advice that finding the investors by analyzing the investment type. Since your friend startup is not in early phase it is in growth stage so the best-suited investment type is Private Equity. Find the top 5 investors who have invested in a different number of startups and their investment type is Private Equity. Correct spelling of investment types are - "Private Equity", "Seed Funding", "Debt Funding", and "Crowd Funding". Keep an eye for any spelling mistake. You can find this by printing unique values from this column.There are many errors in startup names. Ignore correcting all, just handle the important ones - Ola, Flipkart, Oyo and Paytm.

Explanation:

Firstly all columns except "InvestorsName", "InvestmentType" and "StartupName" have been dropped using "drop" function of pandas dataframe (line 8). Then "NA" values of column "InvestorsName", "InvestmentType" and "StartupName" have been drop and finally index are reset using "reset_index" function. Now dataset is ready to get cleaned.

First "InvestmentType" is cleaned and spelling errors have been removed using Boolean array. Then rows which are not required have been dropped. Since, now the startup is not in early stage so we only need "Private Equity" for growth stage startup so "Seed Funding", "Crowd Funding" and "Debt funding" rows have been dropped (line 25) and index are reset. Now StartupName column should be cleaned, only the important startup names like: Ola, Flipkart, Oyo and Paytm has been cleaned using Boolean array.

"startup" variable is used for "StartupName" and "name" variable is used for "InvestorsName" and "split" function is used in order to extract all the names which are separated by commas (line 40,41,42).

Rest other steps are same as in question 3 and 4 (From line 44 onwards).

Finally top 5 investor's name who have invested maximum number of times in different companies having "InvestmenType" as "Seed funding" and "Crowd funding" with their no of investment have been printed and plotted on the graph. X-axis: "investor_name" and Y-axis: "number of investment" of that investor.

Conclusion:

From result it is clear that "Sequoia Capital" has invested most number of times in different companies having "InvestmenType" as "Private Equity". So this investor will be the best for contacting for getting funds and investments for growth stage startup. Then comes "Accel Partners", "Kalaari Capital", "Blume Ventures" and "SAIF Partners".

Thank you.

Abhedya Shukla
9264970112
abhedya21@gmail.com