Analysis of 'startup_funding.csv'

Name-Shubham Ingole.

Que.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.

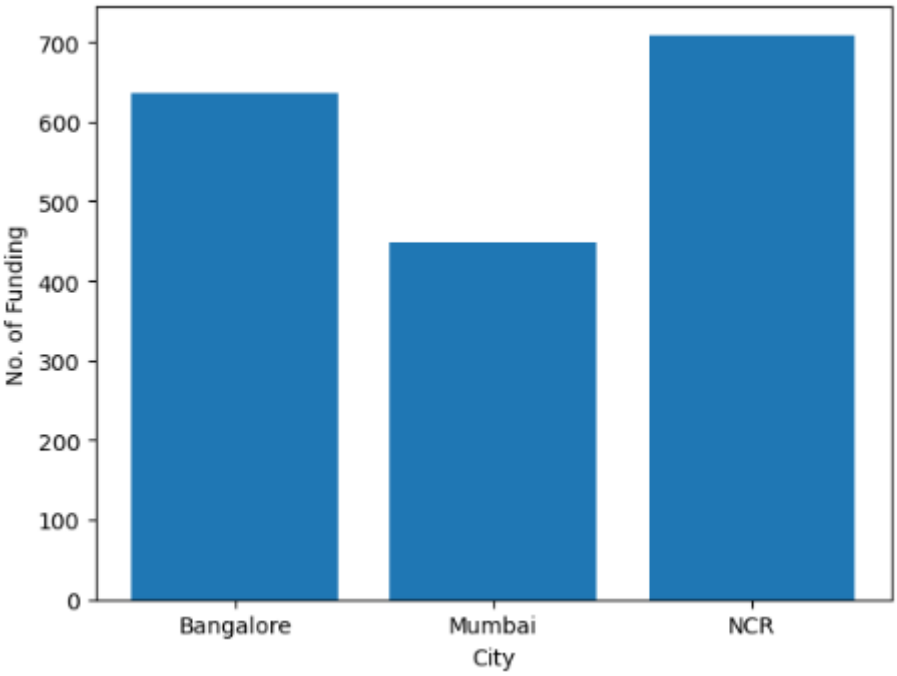
Explanation

1. Importing Required Libraries: pandas, matplotlib and numpy
2. Reading of CSV file named 'startup_funding.csv' and stores the data into a pandas DataFrame called `df`. The 'utf-8' encoding is specified to ensure proper reading of the file.
3. The code performs some data cleaning steps on the 'CityLocation' column. It first dropping any rows with missing values (NaN) in the 'CityLocation' column by using dropna() method. Then, it removes trailing slashes and strips any leading or trailing whitespaces from the strings in the 'CityLocation' column.
4. Replacing wrong city names by with correct one i.e "Delhi" with "New Delhi" and "bangalore" with "Bangalore" in the 'CityLocation' column.
5. Filtering the DataFrame by Specific Cities i.e 'Bangalore', 'Mumbai', 'Gurgaon', 'Noida', 'New Delhi':
6. Counting the Number of Fundings and it is equal to frequency of the city mentioned by using dictionary considering city as key.
7. Printing City with the most Number of Fundings.
8. Creating Arrays using numpy by appending keys of dictionary in list city and appending values in list fre to Plot bar graph.
9. Plotting graph with x-axis having city names and y-axis having no. of funding as frequency

Result

location where the most number of funding is NCR 709

Conclusion



As per the graph we can conclude that location where the most number of funding is NCR with 709 fundings.

Que.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:

1. Importing Required Libraries: pandas, matplotlib and numpy
2. Reading of CSV file named 'startup_funding.csv' and stores the data into a pandas DataFrame called `df`. The 'utf-8' encoding is specified to ensure proper reading of the file.
3. Performing some data cleaning steps on the 'InvestorsName' column. It removes leading and trailing whitespaces from the strings using the `str.strip()` function. It also converts the 'InvestorsName' column to a string data type using `astype(str)`.
4. To ignore undisclosed investor we create empty list as invest and iterating over the 'InvestorsName' column with condition if not `('Undisclosed' in i or 'undisclosed' in i)` to ensure it the name which dosen't contain word undisclosed will be ignored and remaing name will appended in invest list.
5. Creating an empty list `name` and iterates over the cleaned investor names in the `invest` list. It splits each investor name by commas and appends each name segment to the `name` list. This step separates individual investor names if multiple names are listed together.
6. Counting the Number of Fundings and it is equal to the frequency of the investor name in name list mentioned by using dictionary considering investor name as key.
7. Creating empty lists names and times to store the investor names and the corresponding investment counts, respectively. code iterates over the dictionary d, appends each investor name to the names list, and appends the corresponding count to the times list. It then uses NumPy's argsort function to sort the np_times array in descending order and selects the indices of the top 5 investors. Updating the np_name and np_times arrays with the top 5 investor names and their respective counts.

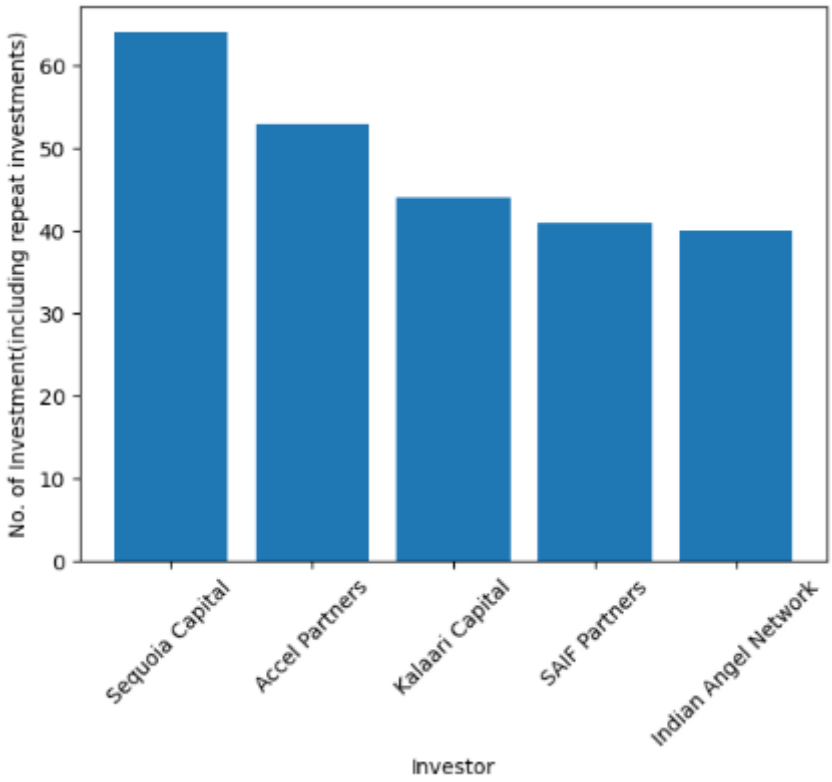
8. Plotting bar chart with x-axis as name of investor and y-axis as no of times investment. x-axis labels are rotated for better readability.

Result:

List of top 5 investors who have invested maximum number of times(considering repeat investments in one company).

(1) Sequoia Capital 64
(2) Accel Partners 53
(3) Kalaari Capital 44
(4) SAIF Partners 41
(5) Indian Angel Network 40

Conclusion:



As per graph we can conclude that 'Sequoia Capital' has the most no of investments i.e 64 time with includin g repetitive investment

---

Que.3) After re-analysing the dataset you found out that some investors have invested in the same startup at dif ferent number of funding rounds. So before finalising the previous list, you want to improvise it by finding the top 5 i nvestors who have invested in different number of startups. This list will be more helpful than your previous list in fi nding 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 co mpany. There are many errors in startup names. Ignore correcting all, just handle the important ones - Ola, Flipkart, Oyo and Paytm.

Explanation:

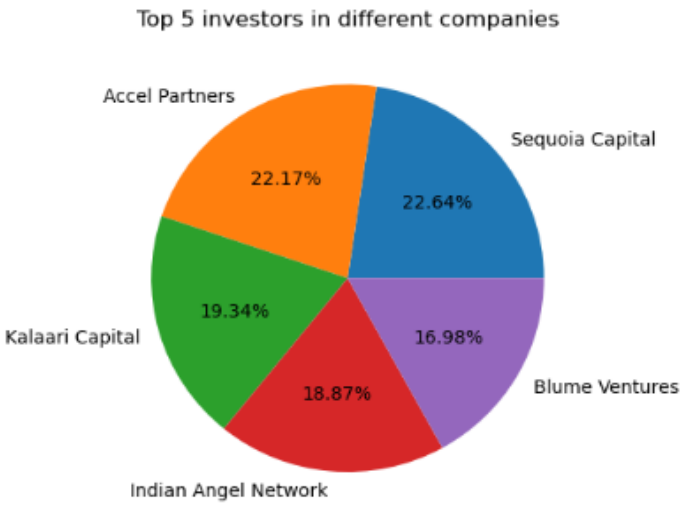
1. Importing Required Libraries: pandas, matplotlib, CSV and numpy
2. Reading of CSV file named 'startup_funding.csv' and stores the data into a pandas DataFrame called `df`. The 'utf-8' encoding is specified to ensure proper reading of the file.
3. Performing several data cleaning steps on the 'InvestorsName' and 'StartupName' columns. It removes leadin g and trailing whitespaces from the 'InvestorsName' column using the `str.strip()` function. It converts th e 'InvestorsName' column to a string data type using `astype(str)`. It replaces variations of the name “Oyo” ,” Paytm”, “Ola”, “Flipkart” with the correct one.
4. Creating empty lists `invest` and `start` to store the investor names and corresponding startup names, respe ctively. Code iterates over the 'InvestorsName' column and will checks if the investor name does not contain the terms "Undisclosed" or "undisclosed". If it the condition is True, it appends the investor name and the co rresponding startup name to the `invest` and `start` lists.
5. Creating empty lists `name` and `startup` to store the unique investor names and their corresponding startup names, respectively. It iterates over the `invest` and `start` lists, splits each investor name by commas, and a ppends each name segment (after stripping leading whitespaces) to the `name` list. It also appends the corresp onding startup name to the `startup` list.
6. Creating a new DataFrame `df` using the `name` and `startup` lists, drops duplicate rows based on both 'Name ' and 'Startup' columns, and removes any rows where the 'Name' column is empty.
7. Creating a dictionary `d` to store the frequency of each investor.
8. Creating empty lists `investor` and `investment` to store the investor names and their respective investme nt counts, respectively. It iterates over the dictionary `d`, appends each investor name to the `investor` list, a nd appends the corresponding count to the `investment` list.
9. By using NumPy's argsort function to sort the np_times array in descending order and selects the indices of t he top 5 investors. Finally, it updates the `np_investor` and `np_investment` arrays with the top 5 investor names and their respective counts.
10. Plotting bar graph as x-axis having np_investor and y-axis having np_investment.

Result

List of top 5 investors who have invested maximum number of times in differen t companies (without repeating).

```
(1) Sequoia Capital 48
(2) Accel Partners 47
(3) Kalaari Capital 41
(4) Indian Angel Network 40
(5) Blume Ventures 36
```

Conclusion



As per graph we can conclude that 'Sequoia Capital' has 22.64% of total investments of top 5 investors i.e 48 investments in different companies.

Que.4) Even after putting so much effort in finding the probable investors, it didn't turn out to be helpful for you r 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 Fundi ng and Crowdfunding. Find the top 5 investors who have invested in a different number of startups and their invest ment type is Crowdfunding or Seed Funding. Correct spelling of investment types are - "Private Equity", "Seed Fundin g", "Debt Funding", and "Crowd Funding". Keep an eye for any spelling mistake. You can find this by printing unique v alues from this column. There are many errors in startup names. Ignore correcting all, just handle the important one s - Ola, Flipkart, Oyo and Paytm.

Explanation:

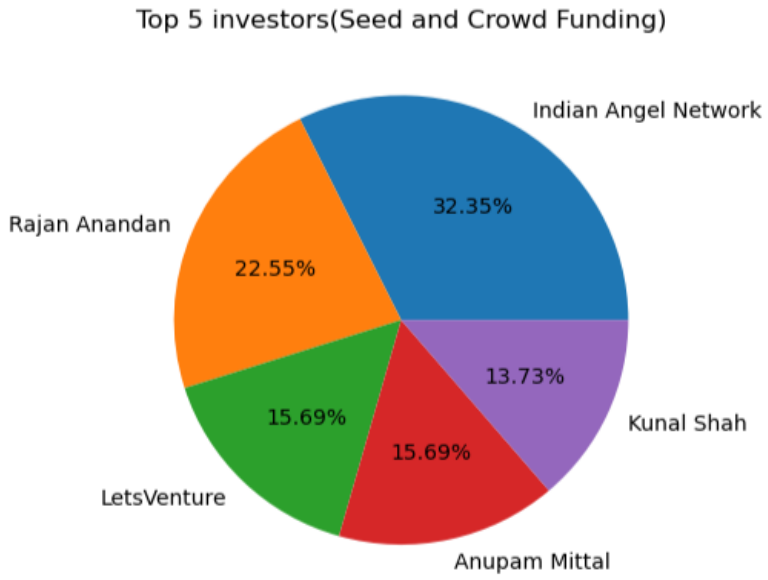
1. Importing Required Libraries: pandas, matplotlib, CSV and numpy
2. Reading of CSV file named 'startup_funding.csv' and stores the data into a pandas DataFrame called `df`. The 'utf-8' encoding is specified to ensure proper reading of the file.
3. Performing datacleaning steps on the 'InvestorsName', 'StartupName', and 'InvestmentType' columns. It rem oves leading and trailing whitespaces from the 'InvestorsName' and 'StartupName' columns using the str.stri p() function. Converting the 'InvestorsName' column to a string data type using astype(str). Performing string replacements to standardize the names of startups ('Oyo', 'Paytm', 'Ola', 'Flipkart') and the investment types ('Private Equity', 'Seed Funding', 'Crowd Funding').
4. Creating empty lists invest, start, and investtype to store the investor names, corresponding startup names, and investment types, respectively. It iterates over the 'InvestorsName' column and checks if the investor na me does not contain the terms "Undisclosed" or "undisclosed". If it passes the condition, it appends the inve stor name, startup name, and investment type to their respective lists.
5. Creating empty lists name, startup, and investmenttype to store the unique investor names, corresponding s tartup names, and investment types, respectively. It iterates over the invest, start, and investtype lists, splits each investor name by commas, and appends each name segment (after stripping leading whitespaces) to th e name list. It also appends the corresponding startup name and investment type to the startup and investm enttype lists.
6. Creating new DataFrame df using the name, startup, and investmenttype lists and drops duplicate rows base d on both 'Name' and 'Startup' columns.
7. As we want only "Crowd Funding" or "Seed Funding".The code filters the DataFrame to retain only the rows where the 'InvestmentType' column values match "Crowd Funding" and "Seed Funding". This step keeps onl y the desired investment types.
8. Dropping the empty value in name column
9. The code creates a dictionary d to store the frequency of each investor.
10. The code creates empty lists investor and investment to store the investor names and their respective invest ment counts, respectively. It iterates over the dictionary d, appends keys of d to the investor list, and value s of d to the investment list
11. By using numpy argsort method to get indices of sorted np_investment in descending order.
12. printed np_investor and np_investment arrays with the top 5 investor names and their respective counts.
13. Plotting bar graph as x-axis having np_investor and y-axis having np_investment.

Result:

List of top 5 investors who have invested in a different number of startups(Investment type is Crowd Funding or Seed Funding)

```
(1) Indian Angel Network 33
(2) Rajan Anandan 23
(3) LetsVenture 16
(4) Anupam Mittal 16
(5) Kunal Shah 14
```

Conclusion:



By seeing graph we can conclude that 'Indian Angel Network' has the most no of investment i.e 33 investments (32.35% of total investment of top 5 investors) in different companies with crowd and seed funding .

Que.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 analysing 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:

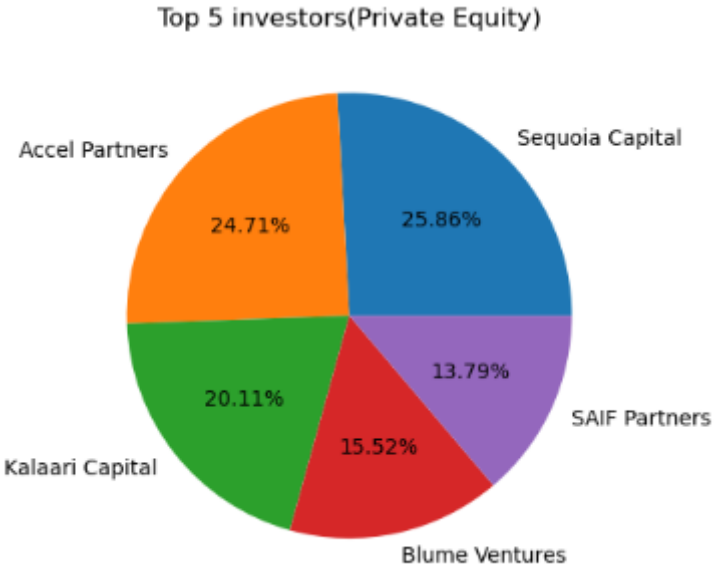
1. The only difference in the 4th problem and 5th is it we have to find the investor who invested max no of times with only 'Private Funding'.
2. Creatd the df by using df=df[df['InvestmentType']=='Private Equity'] this will contains the rows which have 'Private Equity' value in the column 'InvestmentType'
3. Rest of the code is same.
4. Plotting the graph as on x-axis having 'Investor(with Private Equity)' and on y-axis having No. of Investment(in different startup)

Result:

List of top 5 investors who have invested in a different number of startups(Investment type is Private Equity).

(1) Sequoia Capital 45
(2) Accel Partners 43
(3) Kalaari Capital 35
(4) Blume Ventures 27
(5) SAIF Partners 24

Conclusion:



By seeing the graph we can conclude that 'Sequoia Capital' has maximum no of investment i.e 45 investments (25.86 % of total investment of top 5 investors) in different startups with investment type as 'Private Equity'.