Name: Bhavesh Waghela

Student Number: N01639685

Assignment_2

Question_1

In [ ]:
```
Use the link :"https://www.goodreads.com/quotes"
Get the Html file grabs and scrape all the details for each posting.
```

In [1]:
```
import requests
from bs4 import BeautifulSoup

url = "https://www.goodreads.com/quotes"
requests.get(url)
```

Out[1]:
```
<Response [200]>
```

In [2]:
```
page = requests.get(url)

soup = BeautifulSoup(page.text, 'lxml')
soup
```

Out[2]:
```
n._s=r,this.csa=n};

    if (window.csa) {
      window.csa("Config", {
        "Application": "GoodreadsMonolith",
        "Events.SushiEndpoint": "https://unagi.amazon.com/1/events/com.amazon.csm.csa.prod",
        "Events.Namespace": "csa",
        "CacheDetection.RequestID": "REKXWFWDSPWV4574Z885",
        "ObfuscatedMarketplaceId": "A1PQBFHBHS6YH1"
      });

      window.csa("Events")("setEntity", {
        session: { id: "585-4019483-6944928" },
        page: {requestId: "REKXWFWDSPWV4574Z885", meaningful: "interactive"}
      });
    }

    var e = document.createElement("script"); e.src = "https://m.media-amazon.com/images/I/41mrkPcyPwL.js"; document.head.appendChild(e);
  //]]>
```

In [21]:
```
quotes = soup.find_all('div', {'class':'quoteDetails'})
quotes
```

Out[21]:
```
[<div class="quoteDetails">
 <a class="leftAlignedImage quoteAvatar" href="/author/show/3565.Oscar_Wilde">
 <img alt="Oscar Wilde" src="https://i.gr-assets.com/images/S/compressed.photo.goodreads.com/authors/1673611182i/3565._UX200_CR0,15,200,200_.jpg"/>
 </a>
 <div class="quoteText">
       “Be yourself; everyone else is already taken.”
     <br/>
   —
   <span class="authorOrTitle">
     Oscar Wilde
   </span>
 </div>
 <div class="quoteFooter">
 <div class="greyText smallText left">
      tags:
        <a href="/quotes/tag/attributed-no-source">attributed-no-source</a>,
        <a href="/quotes/tag/be-yourself">be-yourself</a>,
        <a href="/quotes/tag/gilbert-perreira">gilbert-perreira</a>,
```

In [32]:
```
extracted_quote = soup.find('div', {'class':'quoteDetails'}).findNext('div').contents[0]
extracted_quote
```

Out[32]:
```
'\n      “Be yourself; everyone else is already taken.”\n    '
```

In [96]:
```
quotes = []
for i in soup.find_all('div', {'class':'quoteDetails'}):
    quote_raw = i.findNext('div').contents[0]
    quote = quote_raw.split('\n      ')[1:]
    quote = ''.join(quote)
    print(quote)
```

```
“Be yourself; everyone else is already taken.”

“I'm selfish, impatient and a little insecure. I make mistakes, I am out of control and at times hard to
handle. But if you can't handle me at my worst, then you sure as hell don't deserve me at my best.”

“Two things are infinite: the universe and human stupidity; and I'm not sure about the universe.”

“So many books, so little time.”

“A room without books is like a body without a soul.”

“Be who you are and say what you feel, because those who mind don't matter, and those who matter don't mi
nd.”

“You've gotta dance like there's nobody watching,
“You know you're in love when you can't fall asleep because reality is finally better than your dreams.”

“You only live once, but if you do it right, once is enough.”

“Be the change that you wish to see in the world.”

“In three words I can sum up everything I've learned about life: it goes on.”

“If you want to know what a man's like, take a good look at how he treats his inferiors, not his equals.”

“Don’t walk in front of me… I may not follow
“If you tell the truth, you don't have to remember anything.”

“Friendship ... is born at the moment when one man says to another "What! You too? I thought that no one
but myself . . .”

“I've learned that people will forget what you said, people will forget what you did, but people will nev
er forget how you made them feel.”

“A friend is someone who knows all about you and still loves you.”

“To live is the rarest thing in the world. Most people exist, that is all.”

“Always forgive your enemies; nothing annoys them so much.”

“Darkness cannot drive out darkness: only light can do that. Hate cannot drive out hate: only love can do
that.”

“Live as if you were to die tomorrow. Learn as if you were to live forever.”

“We accept the love we think we deserve.”

“Without music, life would be a mistake.”

“I am so clever that sometimes I don't understand a single word of what I am saying.”

“To be yourself in a world that is constantly trying to make you something else is the greatest accomplis
hment.”

“Here's to the crazy ones. The misfits. The rebels. The troublemakers. The round pegs in the square hole
s. The ones who see things differently. They're not fond of rules. And they have no respect for the statu
s quo. You can quote them, disagree with them, glorify or vilify them. About the only thing you can't do
is ignore them. Because they change things. They push the human race forward. And while some may see them
as the crazy ones, we see genius. Because the people who are crazy enough to think they can change the wo
rld, are the ones who do.”

“Insanity is doing the same thing, over and over again, but expecting different results.”

“I believe that everything happens for a reason. People change so that you can learn to let go, things go
wrong so that you appreciate them when they're right, you believe lies so you eventually learn to trust n
o one but yourself, and sometimes good things fall apart so better things can fall together.”

“It is better to be hated for what you are than to be loved for what you are not.”

“Twenty years from now you will be more disappointed by the things that you didn't do than by the ones yo
u did do. So throw off the bowlines. Sail away from the safe harbor. Catch the trade winds in your sails.
Explore. Dream. Discover.”
```

Question_2

Analyza the data and answer all questions.

- **How Does the Funding Ecosystem changes with respect to Time?**
- **What is the General Amount that Startups get in India?**
- **Which Kind of Industries are more preferred for Startups?**
- **Does Location also play a role, In determining the Growth of a Startup?**
- **Who plays the main role in Indian Startups Ecosystem?**

- **What are the different Types of Funding for Startups?**

Lets Install the Libraries

In [1]:
```
# install these libraries given below to run the program successfully,
# if already installed ignore.
#pip install numpy
#pip install pandas
#pip seaborn
#pip matplotlib
```

In [155]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Reading the data

In [156]:
```
data = pd.read_csv('startup_funding.csv')

data
```

Out[156]:

	Sr No	Date dd/mm/yyyy	Startup Name	Industry Vertical	SubVertical	City Location	Investors Name	InvestmentnType	Amount i US
0	1	09/01/2020	BYJU'S	E-Tech	E-learning	Bengaluru	Tiger Global Management	Private Equity Round	20,00,00,0(
1	2	13/01/2020	Shuttl	Transportation	App based shuttle service	Gurgaon	Susquehanna Growth Equity	Series C	80,48,3(
2	3	09/01/2020	Mamaearth	E-commerce	Retailer of baby and toddler products	Bengaluru	Sequoia Capital India	Series B	1,83,58,8(
3	4	02/01/2020	https://www.wealthbucket.in/	FinTech	Online Investment	New Delhi	Vinod Khatumal	Pre-series A	30,00,0(
4	5	02/01/2020	Fashor	Fashion and Apparel	Embroiled Clothes For Women	Mumbai	Sprout Venture Partners	Seed Round	18,00,0(
...	...	...	...	...	...	...	...	...	
3038	3040	29/01/2015	Printvenue	NaN	NaN	NaN	Asia Pacific Internet Group	Private Equity	45,00,0(
3039	3041	29/01/2015	Graphene	NaN	NaN	NaN	KARSEMVEN Fund	Private Equity	8,25,0(
3040	3042	30/01/2015	Mad Street Den	NaN	NaN	NaN	Exfinity Fund, GrowX Ventures.	Private Equity	15,00,0(
3041	3043	30/01/2015	Simplotel	NaN	NaN	NaN	MakeMyTrip	Private Equity	Na
3042	3044	31/01/2015	couponmachine.in	NaN	NaN	NaN	UK based Group of Angel Investors	Seed Funding	1,40,0(

3043 rows × 10 columns

In [157]:
```
data = data.rename(columns={'Date dd/mm/yyyy': 'Date', 'City  Location': 'City Location'})
```

In [158]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3043 entries, 0 to 3042
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Sr No              3043 non-null   int64 
 1   Date               3043 non-null   object
 2   Startup Name       3043 non-null   object
 3   Industry Vertical  2872 non-null   object
 4   SubVertical        2108 non-null   object
 5   City Location      2863 non-null   object
 6   Investors Name     3019 non-null   object
 7   InvestmentnType    3039 non-null   object
 8   Amount in USD      2083 non-null   object
 9   Remarks            419 non-null    object
dtypes: int64(1), object(9)
memory usage: 237.9+ KB
```

In [159]:
```
print(data.shape)
data.isna().sum()
```

```
(3043, 10)
```

Out[159]:
```
Sr No                   0
Date                    0
Startup Name            0
Industry Vertical     171
SubVertical           935
City Location         180
Investors Name         24
InvestmentnType         4
Amount in USD         960
Remarks              2624
dtype: int64
```

In [160]:
```
data.duplicated().sum()
```

Out[160]:
```
0
```

In [116]:
```
problematic_dates = data['Date'][pd.to_datetime(data['Date'], format='%d/%m/%Y', errors='coerce').isnull()]
print(problematic_dates)
```

```
2774     12/05.2015
2775     12/05.2015
2830     13/04.2015
3010     15/01.2015
3028    22/01//2015
Name: Date, dtype: object
```

In [161]:
```
problematic_dates = data['Date'][pd.to_datetime(data['Date'], format='%d/%m/%Y', errors='coerce').isnull()]
print(problematic_dates)
```

```
Series([], Name: Date, dtype: object)
```

In [162]:
```
data['Date'] = pd.to_datetime(data['Date'], format='%d/%m/%Y')
```

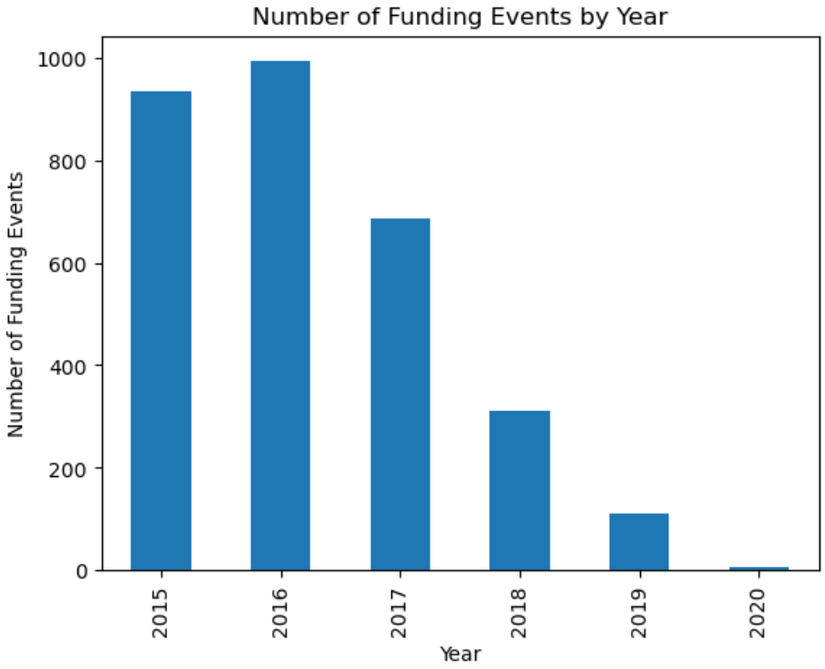
In [163]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3043 entries, 0 to 3042
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype         
---  ------             --------------  -----         
 0   Sr No              3043 non-null   int64         
 1   Date               3043 non-null   datetime64[ns]
 2   Startup Name       3043 non-null   object        
 3   Industry Vertical  2872 non-null   object        
 4   SubVertical        2108 non-null   object        
 5   City Location      2863 non-null   object        
 6   Investors Name     3019 non-null   object        
 7   InvestmentnType    3039 non-null   object        
 8   Amount in USD      2083 non-null   object        
 9   Remarks            419 non-null    object        
dtypes: datetime64[ns](1), int64(1), object(8)
memory usage: 237.9+ KB
```

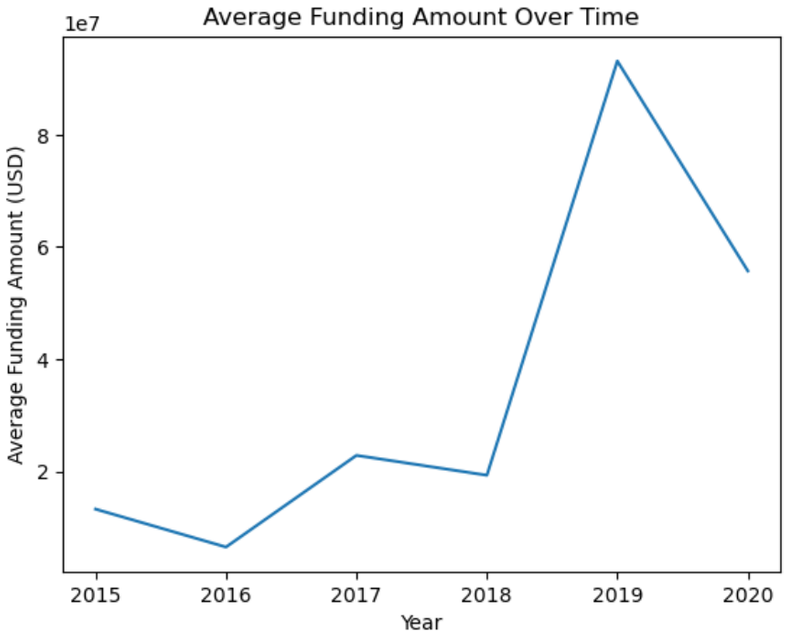
In [ ]:
```

```

How Does the Funding Ecosystem changes with respect to Time?

In [164]:
```
# Explore the Temporal Distribution
data['Year'] = data['Date'].dt.year
data['Month'] = data['Date'].dt.month

data['Year'].value_counts().sort_index().plot(kind='bar')
plt.title('Number of Funding Events by Year')
plt.xlabel('Year')
plt.ylabel('Number of Funding Events')
plt.show()
```



In [165]:
```
# Convert 'Amount in USD' to numeric format
data['Amount in USD'] = pd.to_numeric(data['Amount in USD'].str.replace(',', ''), errors='coerce')

# Plotting the average funding amount over time
data.groupby('Year')['Amount in USD'].mean().plot(kind='line')
plt.title('Average Funding Amount Over Time')
plt.xlabel('Year')
plt.ylabel('Average Funding Amount (USD)')
plt.show()
```

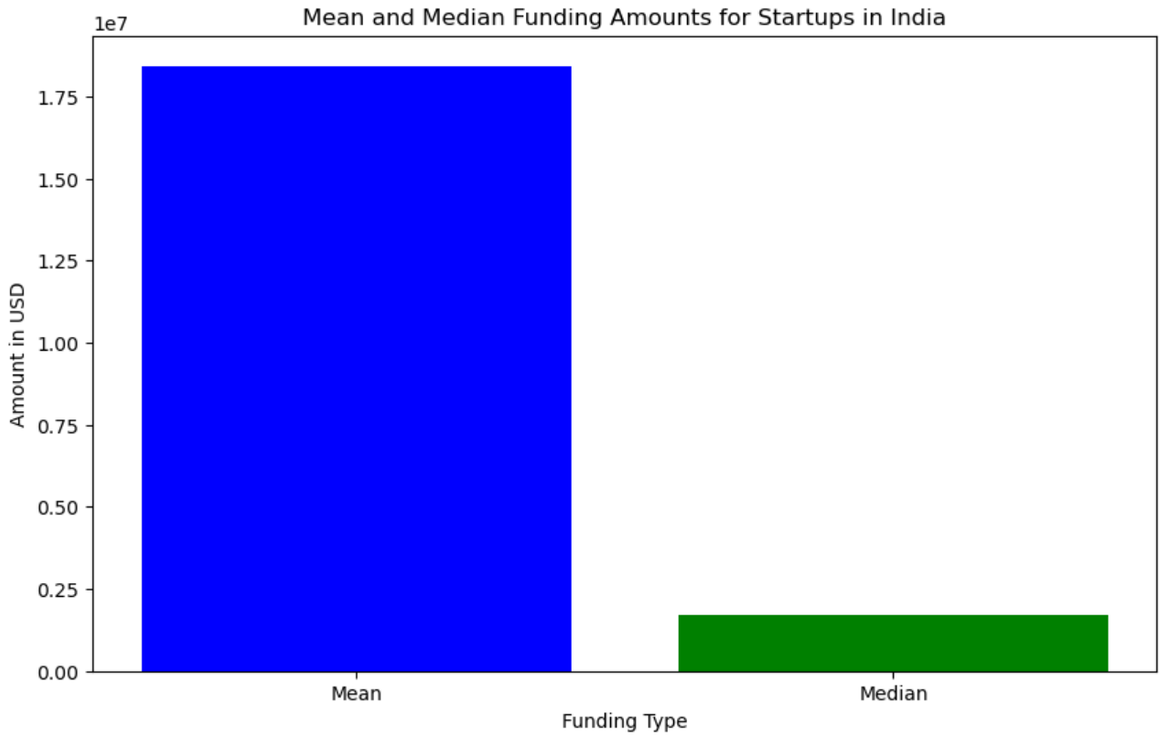


What is the General Amount that Startups get in India?

In [166]:
```
columns_to_clean = ['Startup Name', 'Industry Vertical', 'SubVertical', 'City Location', 'Investors Name',
                    'InvestmentnType', 'Amount in USD', 'Remarks']

# Iterate through each column and replace '\\xc2\\xa0'
for column in columns_to_clean:
    if data[column].dtype == 'O':
        # Use regular expression to replace any occurrences of '\\xc2\\xa0'
        data[column] = data[column].astype(str).replace('\\xc2\\xa0', '')

print(data)
```

```
      Sr No       Date                  Startup Name    Industry Vertical  \
0         1 2020-01-09                       BYJU’S              E-Tech
1         2 2020-01-13                        Shuttl       Transportation
2         3 2020-01-09                     Mamaearth           E-commerce
3         4 2020-01-02  https://www.wealthbucket.in/ (https://www.wealthbucket.in/)              FinTech
4         5 2020-01-02                        Fashor  Fashion and Apparel
...     ...        ...                           ...                 ...
3038   3040 2015-01-29                   Printvenue                 nan
3039   3041 2015-01-29                     Graphene                 nan
3040   3042 2015-01-30                Mad Street Den                nan
3041   3043 2015-01-30                     Simplotel                nan
3042   3044 2015-01-31               couponmachine.in                nan

                              SubVertical City Location  \
0                              E-learning     Bengaluru
1                  App based shuttle service       Gurgaon
2     Retailer of baby and toddler products     Bengaluru
3                          Online Investment     New Delhi
4                Embroiled Clothes For Women        Mumbai
...                                      ...           ...
3038                                   nan           nan
3039                                   nan           nan
3040                                   nan           nan
3041                                   nan           nan
3042                                   nan           nan

                        Investors Name       InvestmentnType  Amount in USD  \
0                Tiger Global Management  Private Equity Round    200000000.0
1              Susquehanna Growth Equity              Series C      8048394.0
2                 Sequoia Capital India              Series B     18358860.0
3                         Vinod Khatumal          Pre-series A      3000000.0
4                Sprout Venture Partners             Seed Round      1800000.0
...                                  ...                   ...           ...
3038        Asia Pacific Internet Group        Private Equity      4500000.0
3039                       KARSEMVEN Fund        Private Equity       825000.0
3040     Exfinity Fund, GrowX Ventures.        Private Equity      1500000.0
3041                           MakeMyTrip        Private Equity            NaN
3042  UK based Group of Angel Investors           Seed Funding       140000.0

                               Remarks  Year  Month
0                                  nan  2020      1
1                                  nan  2020      1
2                                  nan  2020      1
3                                  nan  2020      1
4                                  nan  2020      1
...                                ...   ...    ...
3038                               nan  2015      1
3039               Govt backed VC Fund  2015      1
3040                               nan  2015      1
3041  Strategic Funding, Minority stake  2015      1
3042                               nan  2015      1

[3043 rows x 12 columns]
```

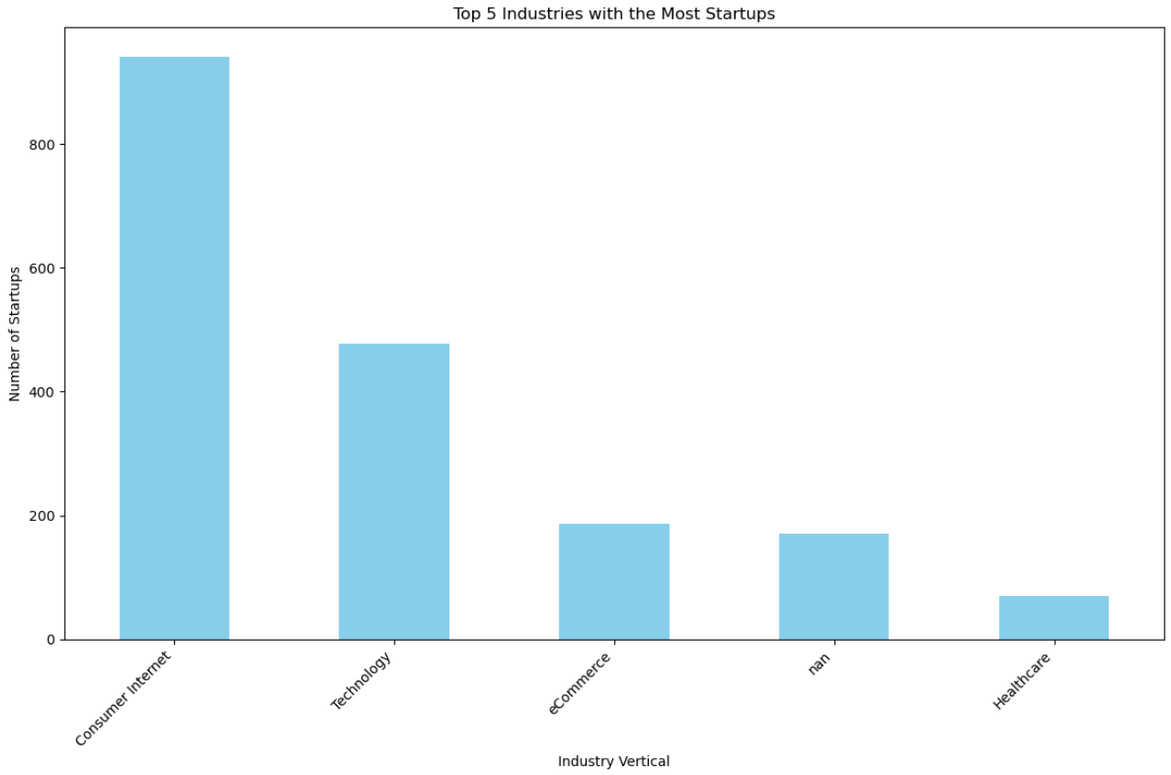
In [167]:
```
# Convert 'Amount in USD' column to numeric, coercing errors to NaN
data['Amount in USD'] = pd.to_numeric(data['Amount in USD'], errors='coerce')

# Filter out rows with NaN values in 'Amount in USD'
funding_data = data.dropna(subset=['Amount in USD'])

# Calculate the mean and median funding amounts
mean_amount = funding_data['Amount in USD'].mean()
median_amount = funding_data['Amount in USD'].median()

# Plotting
plt.figure(figsize=(10, 6))
plt.bar(['Mean', 'Median'], [mean_amount, median_amount], color=['blue', 'green'])
plt.title('Mean and Median Funding Amounts for Startups in India')
plt.xlabel('Funding Type')
plt.ylabel('Amount in USD')
plt.show()
```



In [ ]:

Which Kind of Industries are more preferred for Startups?

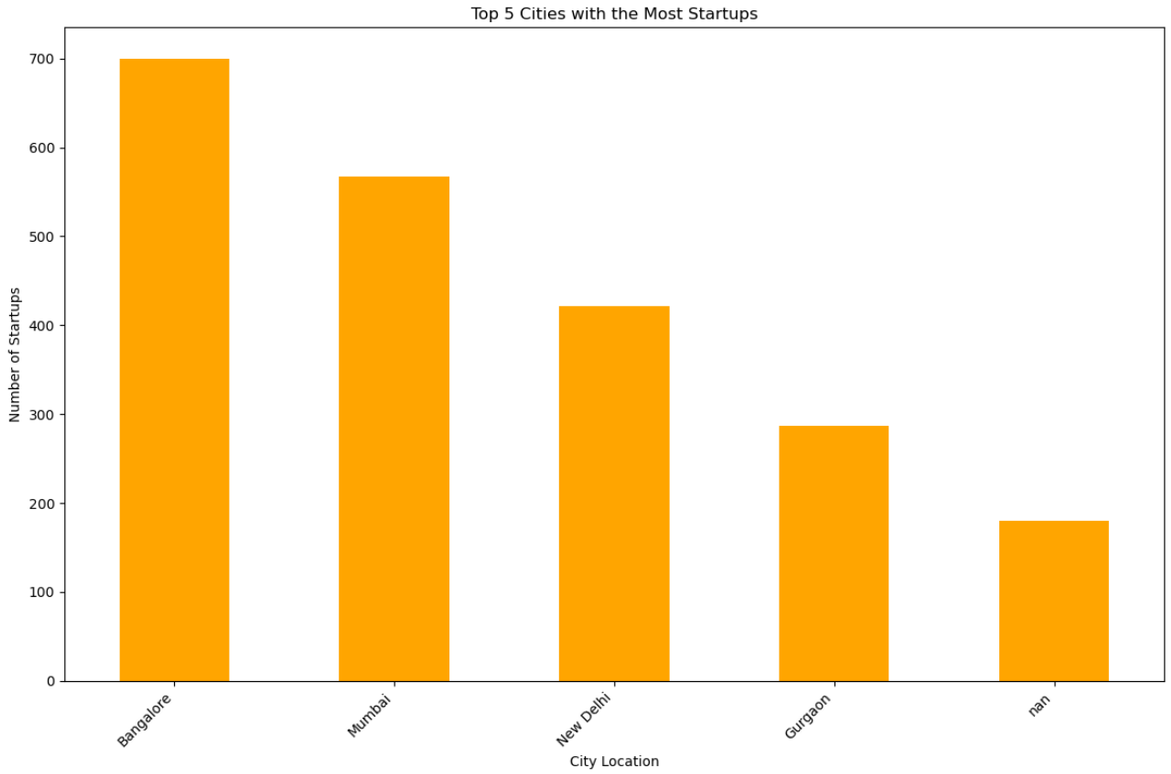
In [168]:
```
# Drop rows with missing values in 'Industry Vertical' column
industry_data = data.dropna(subset=['Industry Vertical'])

# Count the occurrences of each industry
industry_counts = industry_data['Industry Vertical'].value_counts()

# Take only the top 5 industries
top_industries = industry_counts.head(5)

# Plotting
plt.figure(figsize=(12, 8))
top_industries.plot(kind='bar', color='skyblue')
plt.title('Top 5 Industries with the Most Startups')
plt.xlabel('Industry Vertical')
plt.ylabel('Number of Startups')
plt.xticks(rotation=45, ha='right')  # Rotate x-axis labels for better readability
plt.tight_layout()
plt.show()
```



In [ ]:

Does Location also play a role, In determining the Growth of a Startup?

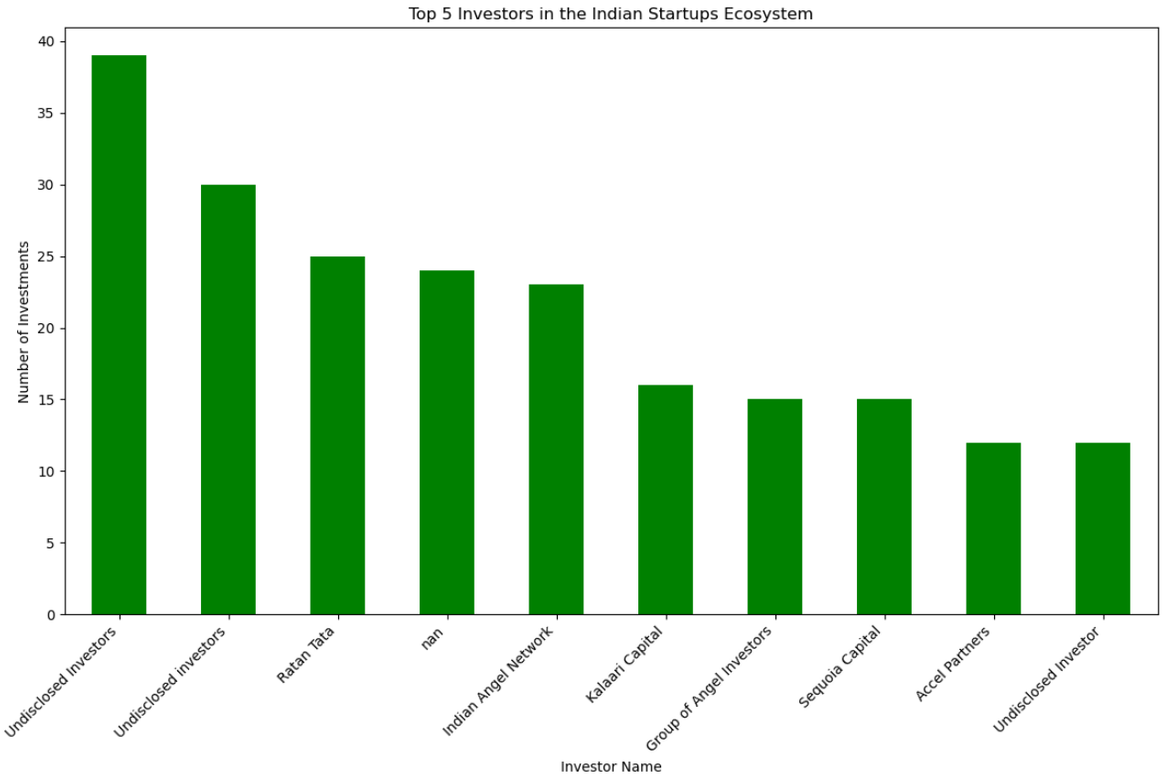
In [170]:
```
# Drop rows with missing values in 'City Location' column
location_data = data.dropna(subset=['City Location'])

# Count the occurrences of each city
location_counts = location_data['City Location'].value_counts()

# Take only the top 5 cities for better visualization
top_locations = location_counts.head(5)

# Plotting
plt.figure(figsize=(12, 8))
top_locations.plot(kind='bar', color='orange')
plt.title('Top 5 Cities with the Most Startups')
plt.xlabel('City Location')
plt.ylabel('Number of Startups')
plt.xticks(rotation=45, ha='right')  # Rotate x-axis labels for better readability
plt.tight_layout()
plt.show()
```



In [ ]:

Who plays the main role in Indian Startups Ecosystem?

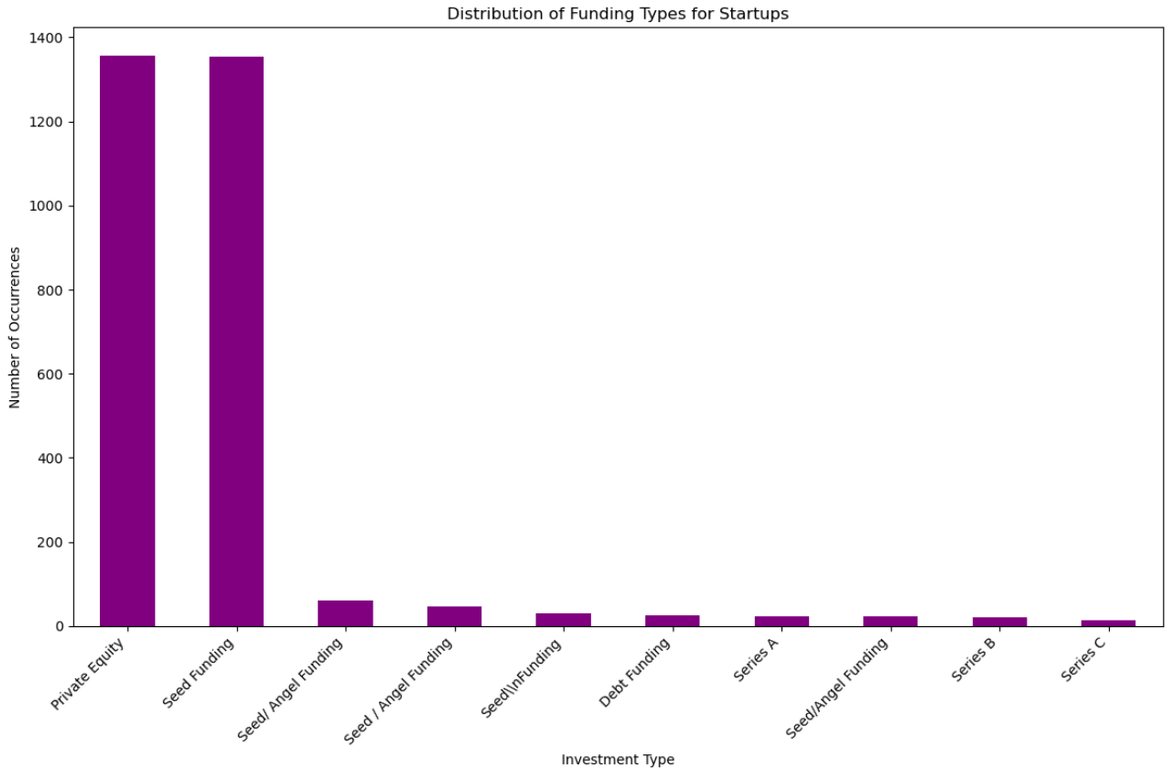
In [172]:
```
# Drop rows with missing values in 'Investors Name' column
investors_data = data.dropna(subset=['Investors Name'])

# Count the occurrences of each investor
investor_counts = investors_data['Investors Name'].value_counts()

# Take only the top 10 investors for better visualization
top_investors = investor_counts.head(10)

# Plotting
plt.figure(figsize=(12, 8))
top_investors.plot(kind='bar', color='green')
plt.title('Top 5 Investors in the Indian Startups Ecosystem')
plt.xlabel('Investor Name')
plt.ylabel('Number of Investments')
plt.xticks(rotation=45, ha='right')  # Rotate x-axis labels for better readability
plt.tight_layout()
plt.show()
```



In [ ]:

What are the different Types of Funding for Startups?

In [175]:
```
# Drop rows with missing values in 'InvestmentnType' column
funding_type_data = data.dropna(subset=['InvestmentnType'])

# Count the occurrences of each funding type
funding_type_counts = funding_type_data['InvestmentnType'].value_counts()

# Take only the top 10 funding type for better visualization
top_funding_type = funding_type_counts.head(10)

# Plotting
plt.figure(figsize=(12, 8))
top_funding_type.plot(kind='bar', color='purple')
plt.title('Distribution of Funding Types for Startups')
plt.xlabel('Investment Type')
plt.ylabel('Number of Occurrences')
plt.xticks(rotation=45, ha='right')  # Rotate x-axis labels for better readability
plt.tight_layout()
plt.show()
```



In [ ]: