```
# This Python 3 environment comes with many helpful analytics
libraries installed
# It is defined by the kaggle/python Docker image:
https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/"
directory
# For example, running this (by clicking run or pressing Shift+Enter)
will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/)
that gets preserved as output when you create a version using "Save &
Run All"
# You can also write temporary files to /kaggle/temp/, but they won't
be saved outside of the current session

import os
import pandas as pd

# List files in the input directory to confirm the correct path
print("Available files:", os.listdir("/kaggle/input/startup-
investments-crunchbase"))

# Load the dataset
file_path =
"/kaggle/input/startup-investments-crunchbase/investments_VC.csv"
df = pd.read_csv(file_path, encoding="latin1")

# Display the first few rows
df.head()
```

```
Available files: ['investments_VC.csv']
```

	permalink	name	homepage_url
0	/organization/waywire	#waywire	http://www.waywire.com
1	/organization/tv-communications	&TV Communications	
2	/organization/rock-your-paper	'Rock' Your Paper	
3	/organization/in-touch-network	(In)Touch Network	
4	/organization/r-ranch-and-mine	-R- Ranch and Mine	

```
1          http://enjoyandtv.com
2   http://www.rockyourpaper.org
3  http://www.InTouchNetwork.com
4                            NaN

                                        category_list        market   \
0         |Entertainment|Politics|Social Media|News|          News
1                                           |Games|         Games
2                             |Publishing|Education|    Publishing
3  |Electronics|Guides|Coffee|Restaurants|Music|i...   Electronics
4                       |Tourism|Entertainment|Games|       Tourism

   funding_total_usd      status country_code state_code
region  ...  \
0          17,50,000    acquired          USA         NY  New York
City  ...
1          40,00,000   operating          USA         CA    Los
Angeles  ...
2             40,000   operating          EST        NaN
Tallinn  ...
3          15,00,000   operating          GBR        NaN
London  ...
4             60,000   operating          USA         TX
Dallas  ...

  secondary_market  product_crowdfunding  round_A  round_B  round_C
round_D  \
0              0.0                   0.0      0.0      0.0      0.0
0.0
1              0.0                   0.0      0.0      0.0      0.0
0.0
2              0.0                   0.0      0.0      0.0      0.0
0.0
3              0.0                   0.0      0.0      0.0      0.0
0.0
4              0.0                   0.0      0.0      0.0      0.0
0.0

  round_E  round_F  round_G  round_H
0     0.0      0.0      0.0      0.0
1     0.0      0.0      0.0      0.0
2     0.0      0.0      0.0      0.0
3     0.0      0.0      0.0      0.0
4     0.0      0.0      0.0      0.0

[5 rows x 39 columns]
```

```python
# Check for missing values in each column
print("Missing Values Per Column:\n", df.isnull().sum())
# Check column data types
```

```python
print("\nColumn Data Types:\n", df.dtypes)
# Check for duplicate rows
print("\nNumber of Duplicate Rows:", df.duplicated().sum())
# Display first 5 rows
df.head()
```

```
Missing Values Per Column:
 permalink                4856
name                      4857
homepage_url              8305
category_list             8817
 market                   8824
 funding_total_usd        4856
status                    6170
country_code             10129
state_code               24133
region                   10129
city                     10972
funding_rounds            4856
founded_at               15740
founded_month            15812
founded_quarter          15812
founded_year             15812
first_funding_at          4856
last_funding_at           4856
seed                      4856
venture                   4856
equity_crowdfunding       4856
undisclosed               4856
convertible_note          4856
debt_financing            4856
angel                     4856
grant                     4856
private_equity            4856
post_ipo_equity           4856
post_ipo_debt             4856
secondary_market          4856
product_crowdfunding      4856
round_A                   4856
round_B                   4856
round_C                   4856
round_D                   4856
round_E                   4856
round_F                   4856
round_G                   4856
round_H                   4856
dtype: int64

Column Data Types:
 permalink                object
```

```
name                      object
homepage_url              object
category_list             object
 market                   object
 funding_total_usd        object
status                    object
country_code              object
state_code                object
region                    object
city                      object
funding_rounds           float64
founded_at                object
founded_month             object
founded_quarter           object
founded_year             float64
first_funding_at          object
last_funding_at           object
seed                     float64
venture                  float64
equity_crowdfunding      float64
undisclosed              float64
convertible_note         float64
debt_financing           float64
angel                    float64
grant                    float64
private_equity           float64
post_ipo_equity          float64
post_ipo_debt            float64
secondary_market         float64
product_crowdfunding     float64
round_A                  float64
round_B                  float64
round_C                  float64
round_D                  float64
round_E                  float64
round_F                  float64
round_G                  float64
round_H                  float64
dtype: object

Number of Duplicate Rows: 4855

                        permalink                name  \
0            /organization/waywire            #waywire
1  /organization/tv-communications  &TV Communications
2    /organization/rock-your-paper   'Rock' Your Paper
3   /organization/in-touch-network   (In)Touch Network
4   /organization/r-ranch-and-mine  -R- Ranch and Mine

                     homepage_url  \
```

```
0         http://www.waywire.com
1          http://enjoyandtv.com
2   http://www.rockyourpaper.org
3  http://www.InTouchNetwork.com
4                            NaN

                                        category_list        market   \
0         |Entertainment|Politics|Social Media|News|          News
1                                          |Games|         Games
2                            |Publishing|Education|    Publishing
3  |Electronics|Guides|Coffee|Restaurants|Music|i...   Electronics
4                       |Tourism|Entertainment|Games|       Tourism

   funding_total_usd      status country_code state_code
region  ...  \
0          17,50,000    acquired          USA         NY  New York
City  ...
1          40,00,000   operating          USA         CA    Los
Angeles  ...
2             40,000   operating          EST        NaN
Tallinn  ...
3          15,00,000   operating          GBR        NaN
London  ...
4             60,000   operating          USA         TX
Dallas  ...

  secondary_market  product_crowdfunding round_A round_B round_C
round_D  \
0              0.0                   0.0     0.0     0.0     0.0
0.0
1              0.0                   0.0     0.0     0.0     0.0
0.0
2              0.0                   0.0     0.0     0.0     0.0
0.0
3              0.0                   0.0     0.0     0.0     0.0
0.0
4              0.0                   0.0     0.0     0.0     0.0
0.0

  round_E round_F  round_G  round_H
0     0.0     0.0      0.0      0.0
1     0.0     0.0      0.0      0.0
2     0.0     0.0      0.0      0.0
3     0.0     0.0      0.0      0.0
4     0.0     0.0      0.0      0.0

[5 rows x 39 columns]
```

```
print(df.columns.tolist())
```

```
['permalink', 'name', 'homepage_url', 'category_list', ' market ', '
funding_total_usd ', 'status', 'country_code', 'city',
'funding_rounds', 'founded_at', 'first_funding_at', 'last_funding_at',
'seed', 'venture', 'equity_crowdfunding', 'undisclosed',
'convertible_note', 'debt_financing', 'angel', 'grant',
'private_equity', 'post_ipo_equity', 'post_ipo_debt',
'secondary_market', 'product_crowdfunding', 'round_A', 'round_B',
'round_C', 'round_D', 'round_E', 'round_F', 'round_G', 'round_H']
```

Data Cleaning

```
# Remove leading and trailing spaces from all column names
df.columns = df.columns.str.strip()

# Print the updated column names to verify the change
print(df.columns.tolist())
```

```
['permalink', 'name', 'homepage_url', 'category_list', 'market',
'funding_total_usd', 'status', 'country_code', 'city',
'funding_rounds', 'founded_at', 'first_funding_at', 'last_funding_at',
'seed', 'venture', 'equity_crowdfunding', 'undisclosed',
'convertible_note', 'debt_financing', 'angel', 'grant',
'private_equity', 'post_ipo_equity', 'post_ipo_debt',
'secondary_market', 'product_crowdfunding', 'round_A', 'round_B',
'round_C', 'round_D', 'round_E', 'round_F', 'round_G', 'round_H']
```

```
# Check for missing values in each column again
print(df.isnull().sum())
```

```
permalink                   1
name                        2
homepage_url             3450
category_list            3962
market                   3969
funding_total_usd           1
status                   1315
country_code             5274
city                     6117
funding_rounds              1
founded_at              10885
first_funding_at            1
last_funding_at             1
seed                        1
venture                     1
equity_crowdfunding         1
undisclosed                 1
convertible_note            1
debt_financing              1
angel                       1
grant                       1
```

```
private_equity                1
post_ipo_equity               1
post_ipo_debt                 1
secondary_market              1
product_crowdfunding          1
round_A                       1
round_B                       1
round_C                       1
round_D                       1
round_E                       1
round_F                       1
round_G                       1
round_H                       1
dtype: int64
```

```
# Drop rows where critical columns are missing (name, category,
funding amount)
df = df.dropna(subset=["name", "category_list", "funding_total_usd",
"market"])

# Check missing values again
print(df.isnull().sum())
```

```
permalink                  0
name                       0
homepage_url            2394
category_list              0
market                     0
funding_total_usd          0
status                   935
country_code            4140
city                    4887
funding_rounds             0
founded_at              8910
first_funding_at           0
last_funding_at            0
seed                       0
venture                    0
equity_crowdfunding        0
undisclosed                0
convertible_note           0
debt_financing             0
angel                      0
grant                      0
private_equity             0
post_ipo_equity            0
post_ipo_debt              0
secondary_market           0
product_crowdfunding       0
round_A                    0
```

```
round_B                     0
round_C                     0
round_D                     0
round_E                     0
round_F                     0
round_G                     0
round_H                     0
dtype: int64
```

```
# Fill missing categorical values with "Unknown"
df["status"] = df["status"].fillna("Unknown")
df["country_code"] = df["country_code"].fillna("Unknown")
df["city"] = df["city"].fillna("Unknown")

# Check missing values again
print(df.isnull().sum())
```

```
permalink                   0
name                        0
homepage_url                0
category_list               0
market                      0
funding_total_usd        7071
status                      0
country_code                0
city                        0
funding_rounds              0
founded_at                  0
first_funding_at            9
last_funding_at             0
seed                        0
venture                     0
equity_crowdfunding         0
undisclosed                 0
convertible_note            0
debt_financing              0
angel                       0
grant                       0
private_equity              0
post_ipo_equity             0
post_ipo_debt               0
secondary_market            0
product_crowdfunding        0
round_A                     0
round_B                     0
round_C                     0
round_D                     0
round_E                     0
round_F                     0
round_G                     0
```

```
round_H                    0
funding_year               9
dtype: int64

# Fill missing homepage URLs with "No Website"
df["homepage_url"] = df["homepage_url"].fillna("No Website")

# Check missing values again
print(df.isnull().sum())
```

```
permalink                  0
name                       0
homepage_url               0
category_list              0
market                     0
funding_total_usd       7071
status                     0
country_code               0
city                       0
funding_rounds             0
founded_at                 0
first_funding_at           9
last_funding_at            0
seed                       0
venture                    0
equity_crowdfunding        0
undisclosed                0
convertible_note           0
debt_financing             0
angel                      0
grant                      0
private_equity             0
post_ipo_equity            0
post_ipo_debt              0
secondary_market           0
product_crowdfunding       0
round_A                    0
round_B                    0
round_C                    0
round_D                    0
round_E                    0
round_F                    0
round_G                    0
round_H                    0
funding_year               9
dtype: int64
```

```
df["founded_at"] = df["founded_at"].fillna("Unknown")
```

```python
# Check missing values again
print(df.isnull().sum())
```

```
permalink                  0
name                       0
homepage_url               0
category_list              0
market                     0
funding_total_usd       7071
status                     0
country_code               0
city                       0
funding_rounds             0
founded_at                 0
first_funding_at           9
last_funding_at            0
seed                       0
venture                    0
equity_crowdfunding        0
undisclosed                0
convertible_note           0
debt_financing             0
angel                      0
grant                      0
private_equity             0
post_ipo_equity            0
post_ipo_debt              0
secondary_market           0
product_crowdfunding       0
round_A                    0
round_B                    0
round_C                    0
round_D                    0
round_E                    0
round_F                    0
round_G                    0
round_H                    0
funding_year               9
dtype: int64
```

```python
# Find unique values that are not numbers
print(df["funding_total_usd"].unique()[:20])  # Show the first 20
unique values
```

```
[' 17,50,000 ' ' 40,00,000 ' ' 40,000 ' ' 15,00,000 ' ' 60,000 '
 ' 70,00,000 ' ' 49,12,393 ' ' 20,00,000 ' ' -    ' ' 41,250 '
 ' 44,00,000 ' ' 20,50,000 ' ' 5,00,000 ' ' 25,35,000 ' ' 49,62,651 '
 ' 40,59,079 ' ' 1,00,00,000 ' ' 30,00,000 ' ' 45,00,000 ' ' 4,20,000
']
```

```
# Replace non-numeric values (" - " or empty spaces) with NaN
df["funding_total_usd"] = df["funding_total_usd"].replace([" - ", "",
"unknown"], float("nan"))

# Now safely convert to float
df["funding_total_usd"] = df["funding_total_usd"].replace(",", "",
regex=True).astype(float)

# Confirm the conversion
print(df["funding_total_usd"].dtype)
```

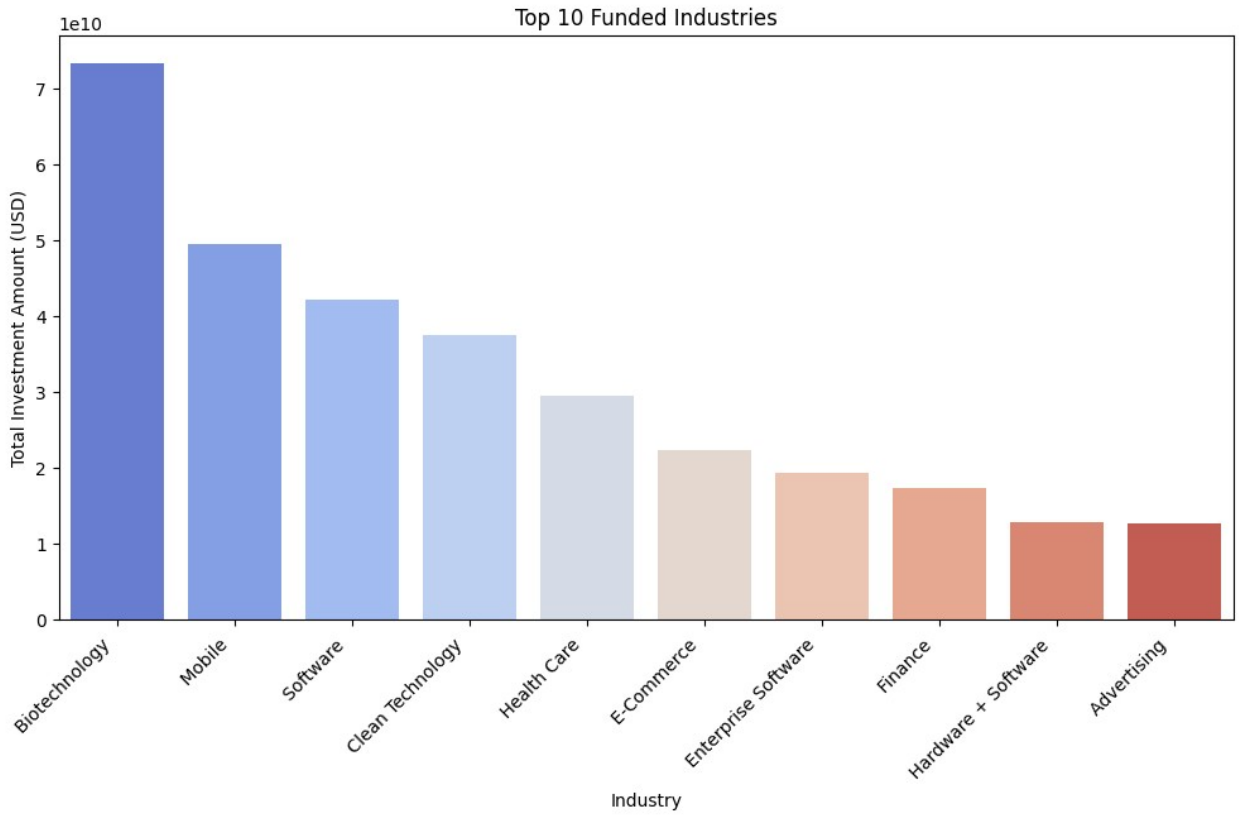
```
float64
```

Exploratory Data Analysis (EDA)

```
import matplotlib.pyplot as plt
import seaborn as sns

# Group by Industry and sum funding amounts
industry_funding = df.groupby("market")
["funding_total_usd"].sum().sort_values(ascending=False).head(10)

# Plot
plt.figure(figsize=(12, 6))
sns.barplot(x=industry_funding.index, y=industry_funding.values,
palette="coolwarm")
plt.xticks(rotation=45, ha='right')
plt.xlabel("Industry")
plt.ylabel("Total Investment Amount (USD)")
plt.title("Top 10 Funded Industries")
plt.show()
```

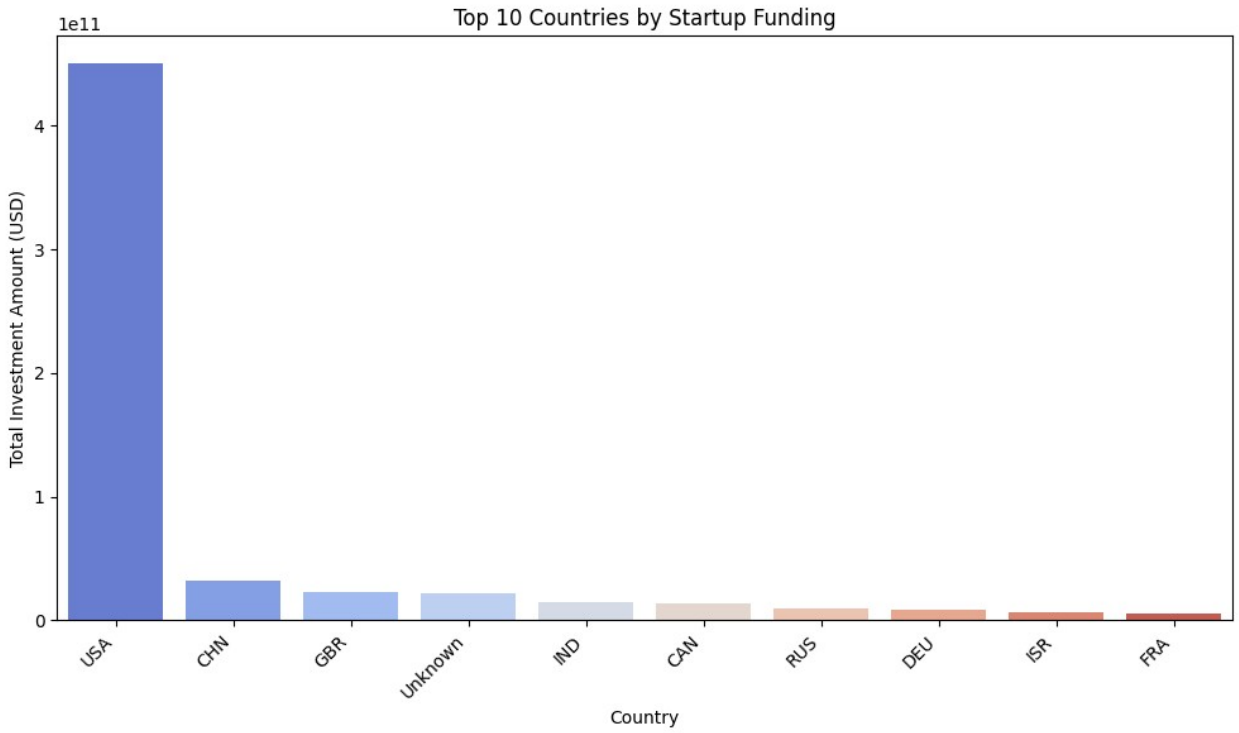


This graph illustrates the total investment across various industries, showcasing which sectors attract the most venture capital. Biotechnology leads the investment landscape, reflecting the high costs and potential breakthroughs in medical research and pharmaceutical innovation. Mobile and Software sectors follow closely, driven by the proliferation of smartphones, cloud computing, and AI applications. The HealthTech and Clean Technology industries also receive substantial funding, highlighting the growing emphasis on healthcare advancements and sustainable solutions. The relatively even distribution among the top industries suggests that venture capital is diversifying across different technology-driven markets, rather than being concentrated in a single sector.

Further Analysis

```
# Group by Country and sum funding amounts
country_funding = df.groupby("country_code")
["funding_total_usd"].sum().sort_values(ascending=False).head(10)

# Plot
plt.figure(figsize=(12, 6))
sns.barplot(x=country_funding.index, y=country_funding.values,
palette="coolwarm")
plt.xticks(rotation=45, ha='right')
plt.xlabel("Country")
plt.ylabel("Total Investment Amount (USD)")
plt.title("Top 10 Countries by Startup Funding")
plt.show()
```



This visualization highlights the geographical distribution of startup funding, with the United States overwhelmingly leading the market. The dominance of the USA can be attributed to Silicon Valley, a highly developed VC ecosystem, and a culture of innovation. China follows at a distant second, driven by government-backed tech investments, AI, and FinTech growth. The UK, India, and Canada also receive significant investments, reinforcing their status as emerging tech hubs with strong startup ecosystems. Interestingly, the presence of an "Unknown" category suggests data inconsistencies or startups that operate across multiple regions without clear country classification.

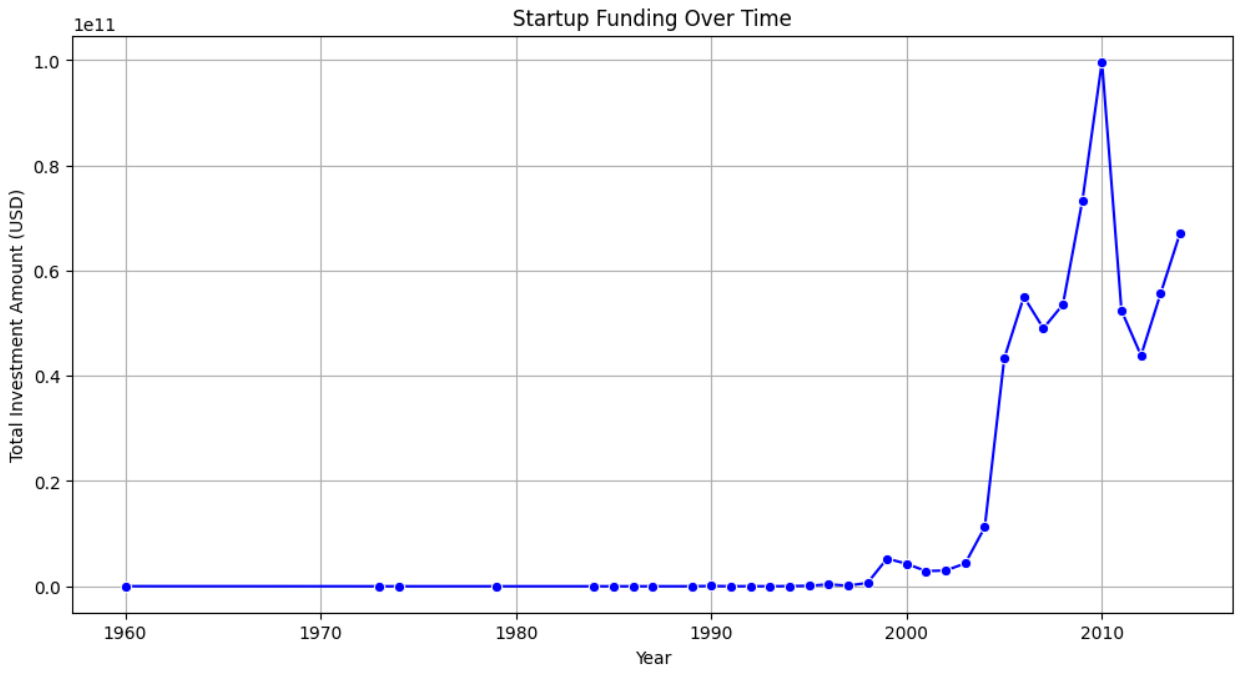
```
# Convert 'first_funding_at' to datetime format
df["first_funding_at"] = pd.to_datetime(df["first_funding_at"],
errors="coerce")

# Extract the year from 'first_funding_at'
df["funding_year"] = df["first_funding_at"].dt.year

# Group by year and sum funding amounts
yearly_funding = df.groupby("funding_year")["funding_total_usd"].sum()

# Plot
plt.figure(figsize=(12, 6))
sns.lineplot(x=yearly_funding.index, y=yearly_funding.values,
marker="o", color="b")
plt.xlabel("Year")
plt.ylabel("Total Investment Amount (USD)")
plt.title("Startup Funding Over Time")
```

```
plt.grid(True)
plt.show()
```



The trend over time reveals a slow but steady increase in funding until the late 1990s, after which there is a significant jump. The early 2000s funding surge aligns with the dot-com boom, while the subsequent dip corresponds to the dot-com bust. Another massive funding increase is seen post-2005, coinciding with the rise of social media, cloud computing, and AI-driven companies. The peak around 2010 suggests a high-investment period, likely due to mobile technology expansion, deep learning breakthroughs, and the growth of unicorn startups like Uber and Airbnb. The fluctuations after 2015 indicate changing investment trends, economic cycles, and shifts in VC preferences.

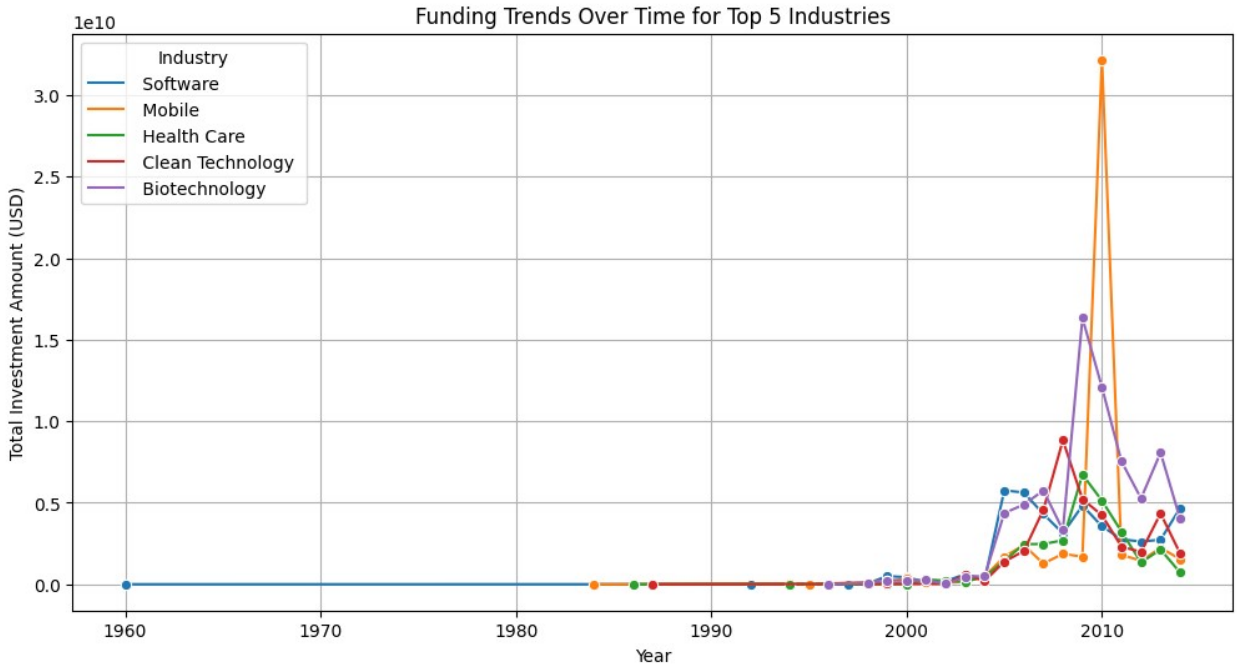
```
import matplotlib.pyplot as plt
import seaborn as sns

# Group by industry and year to see trends
industry_yearly_funding = df.groupby(["funding_year", "market"])
["funding_total_usd"].sum().reset_index()

# Filter for the top 5 industries
top_industries = df.groupby("market")
["funding_total_usd"].sum().sort_values(ascending=False).head(5).index
industry_yearly_funding =
industry_yearly_funding[industry_yearly_funding["market"].isin(top_ind
ustries)]

# Plot
plt.figure(figsize=(12, 6))
```

```
sns.lineplot(data=industry_yearly_funding, x="funding_year",
y="funding_total_usd", hue="market", marker="o")
plt.xlabel("Year")
plt.ylabel("Total Investment Amount (USD)")
plt.title("Funding Trends Over Time for Top 5 Industries")
plt.legend(title="Industry")
plt.grid(True)
plt.show()
```



This graph provides a deeper look into how funding patterns evolved across key industries. Software and Mobile saw rapid growth in the 2000s, with Mobile peaking significantly around 2010, likely due to the rise of smartphones and app-based businesses. Biotechnology and HealthTech funding show a more stable increase, reflecting ongoing interest in medical innovations, personalized healthcare, and biotech advancements. Clean Technology, though growing steadily, appears more volatile, potentially due to government regulations, sustainability trends, and shifts in renewable energy investments. The diversity in funding trajectories suggests that different industries experience investment waves based on technological breakthroughs, market needs, and regulatory environments.

Conclusion: Key Insights from Startup Funding Analysis

In this project, we explored startup funding trends using a dataset of venture investments across industries, countries, and time periods. Here are the main takeaways from our analysis:

Industries with the Most Funding

- Biotechnology received the highest total investment, followed by Mobile, Software, Clean Technology, and Healthcare.

- This indicates a strong investor interest in cutting-edge medical advancements and tech-driven industries.

Countries Dominating Startup Investments

- The USA overwhelmingly leads in startup funding, followed by China, the UK, and India.
- This suggests that startup ecosystems in these countries attract significant venture capital activity.

Trends in Startup Funding Over Time

- There was a massive surge in investments post-2000, with funding peaking between 2008 and 2015.
- This aligns with major tech booms, increased access to venture capital, and a rise in AI, SaaS, and FinTech startups.