CSM-353: EDA Project

Analysis of Startup Investments in India

B.Tech CSE

(Data Science with ML)

Submitted to

Mr. Ved Prakash Chaubey

LOVELY PROFESSIONAL UNIVERSITY

PHAGWARA, PUNJAB

SUBMITTED BY

Name of student : Zubair Ahmed P

Registration Number : 12201412

DECLARATION

I Zubair Ahmed Proddutur, hereby declare that the work done by me on "Analysis of Startup Investments in India" from September 2024 to November 2024, is a record of original work for the partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Computer Science - Data Science with ML, Lovely Professional University, Phagwara.

Student Signature

Zubair Ahmed P

12201412

faculty Signature

Mr. Ved Prakash Chaubey

UID: 63892

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my University and UpGrad for providing me with the golden opportunity to work on this exciting project in Exploratory Data Analysis. This project has not only broadened my understanding of data science but also significantly contributed to my professional growth by giving me hands-on experience in analyzing real-world datasets.

I would like to extend my heartfelt thanks to all those who supported me in developing the 'Analysis of Startup Investments in India' project. Their guidance and encouragement were invaluable throughout the process

I would also like to express my deep appreciation to my mentor,

Mr. Ved Prakash Chaubey, for his invaluable advice, feedback, and guidance throughout the project. His insights and expertise were critical in refining the project and enhancing its overall performance. Without his dedicated support, this project would not have been possible.

Table of Content

Abstract

This project delves into a comprehensive analysis of startup funding in India to uncover trends, patterns, and unique insights into the dynamic Indian startup ecosystem. The dataset encompasses crucial details such as funding amounts, startup locations, investment types, and industry verticals.

Through data cleaning and preprocessing, missing values and inconsistencies were addressed, ensuring the reliability of subsequent analyses. By Implementing Exploratory Data Analysis (EDA), the study reveals valuable insights, including the geographical distribution of startups, funding seasonality, investor diversity across industries, and trends in funding over time.

Unique visualizations—such as co-investor network analysis, funding seasonality patterns, and word clouds of startup names—offer novel perspectives on investor behavior and industry trends. Additionally, the project emphasizes the interplay between funding amounts and city-specific startup ecosystems, as well as the evolution of key industries like EdTech.

This analysis not only provides a snapshot of the Indian startup landscape but also serves as a foundation for future research and decision-making for entrepreneurs, investors, and policymakers.

Introduction

The Indian startup ecosystem has witnessed exponential growth, attracting global attention and significant funding. This project leverages a

dataset comprising 3,044 records and 10 key attributes to analyze trends and patterns in startup funding across India. The dataset provides comprehensive details, including startup names, industry verticals, investment types, funding amounts, investors, city locations, and funding dates.

The primary objective of this analysis is to uncover actionable insights that illuminate the dynamics of the Indian startup landscape. Through exploratory data analysis (EDA), this study aims to:

1. Identify startups that have received the highest funding.
2. Examine the geographical distribution of startups across India.
3. Analyze funding trends over time to understand growth trajectories.
4. Investigate the distribution of investment types and industries attracting the most capital.

Features of the Dataset:

- Date (dd/mm/yyyy): The date when funding was received.
- Startup Name: The name of the startup receiving funding.
- Industry Vertical: The industry classification (e.g., E-commerce, FinTech).
- SubVertical: A more specific industry subcategory (e.g., Agritech, Online Learning).
- City Location: The city where the startup is based.
- Investors Name: The names of investors involved in funding rounds.
- Investment Type: The type of investment round (e.g., Seed, Series A, Private Equity).
- Amount in USD: The funding amount received, requiring data cleaning due to formatting inconsistencies.
- Remarks: Additional notes on funding, which are largely missing in the dataset.

Methodology

Data Collection and Loading:

- The dataset startup_funding.csv was loaded using Pandas to ensure efficient handling of tabular data.
- Initial exploration was conducted using head() and info() functions to understand the structure and data types.

Data Preprocessing:

- Date Conversion: Converted the Date dd/mm/yyyy column to a proper datetime format for temporal analysis.
- Handling Amount Data: Cleaned the Amount in USD column by removing non-numeric characters (e.g., commas) and converted it to a numeric data type.
- Missing Values:
    - Dropped rows with missing Date dd/mm/yyyy.
    - Filled missing values in columns like Industry Vertical, SubVertical, City Location, Investors Name, and InvestmentnType with 'Unknown'.

    - Replaced missing funding amounts with 0.
- Irrelevant Columns: Removed the Remarks column as it was not relevant to the analysis.

Outlier Removal:

- Calculated the Interquartile Range (IQR) to identify outliers in Amount in USD.
- Removed rows with funding amounts beyond 1.5 times the IQR for data consistency.

Feature Engineering:

- Extracted Year and Month from the date column for trend analysis.
- Standardized City Location and Industry Vertical values to ensure uniformity (e.g., converting 'Bengaluru' to 'Bangalore').

Exploratory Data Analysis (EDA):

- Temporal Trends:
    - Analyzed yearly and monthly funding trends using line plots.
    - Grouped data by Year to observe trends in total funding and deal counts.
- City and Industry Insights:
    - Aggregated funding data by City Location and Industry Vertical to identify top-performing regions and sectors.
- Investment Type Analysis:
    - Grouped data by Investment Type to explore the most common funding mechanisms.

- Distribution Analysis:
    - Visualized funding distributions using histograms.
    - Used bar plots to identify top cities and industries by funding and startup count.
- City-Industry Matrix:
    - Created a heatmap to represent average funding amounts for city-industry pairs.

Co-Investor Network Analysis:

- Cleaned and split investor names.
- Generated co-investor combinations to identify top partnerships.
- Visualized co-investor relationships using bar plots.

Visualization:

- Used matplotlib and seaborn for creating:
    - Line plots for temporal trends.
    - Bar plots for city, industry, and investment-type comparisons.
    - Heatmaps for city-industry funding relationships.
    - Pair plots to analyze correlations between continuous variables.

Advanced Analysis:

- Analyzed funding trends over time for specific industries using multi-line plots.
- Explored startup counts across years for top industries.

Tools and Libraries:

- Pandas: For data manipulation and preprocessing.
- NumPy: For numerical operations and handling missing values.
- Matplotlib & Seaborn: For creating insightful visualizations.

Code Analysis & Visual Representations

Importing libraries

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

```
[ ] startup_data = pd.read_csv('startup_funding.csv')
    startup_data.head(10)
```

	Sr No	Date dd/mm/yyyy	Startup Name	Industry Vertical	SubVertical	City Location	Investors Name	InvestmentnType	Amount in USD	Remarks
0	1	09/01/2020	BYJU'S	E-Tech	E-learning	Bengaluru	Tiger Global Management	Private Equity Round	20,00,00,000	NaN
1	2	13/01/2020	Shuttl	Transportation	App based shuttle service	Gurgaon	Susquehanna Growth Equity	Series C	80,48,394	NaN
2	3	09/01/2020	Mamaearth	E-commerce	Retailer of baby and toddler products	Bengaluru	Sequoia Capital India	Series B	1,83,58,860	NaN
3	4	02/01/2020	https://www.wealthbucket.in/	FinTech	Online Investment	New Delhi	Vinod Khatumal	Pre-series A	30,00,000	NaN
4	5	02/01/2020	Fashor	Fashion and Apparel	Embroiled Clothes For Women	Mumbai	Sprout Venture Partners	Seed Round	18,00,000	NaN
5	6	13/01/2020	Pando	Logistics	Open-market, freight management platform	Chennai	Chiratae Ventures	Series A	90,00,000	NaN
6	7	10/01/2020	Zomato	Hospitality	Online Food Delivery Platform	Gurgaon	Ant Financial	Private Equity Round	15,00,00,000	NaN
7	8	12/12/2019	Ecozen	Technology	Agritech	Pune	Sathguru Catalyzer Advisors	Series A	60,00,000	NaN
8	9	06/12/2019	CarDekho	E-Commerce	Automobile	Gurgaon	Ping An Global Voyager Fund	Series D	7,00,00,000	NaN
9	10	03/12/2019	Dhruva Space	Aerospace	Satellite Communication	Bengaluru	Mumbai Angels, Ravikanth Reddy	Seed	5,00,00,000	NaN

```
startup_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3044 entries, 0 to 3043
```

```
[ ] startup_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3044 entries, 0 to 3043
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Sr No              3044 non-null   int64
 1   Date dd/mm/yyyy    3044 non-null   object
 2   Startup Name       3044 non-null   object
 3   Industry Vertical  2873 non-null   object
 4   SubVertical        2108 non-null   object
 5   City  Location     2864 non-null   object
 6   Investors Name     3020 non-null   object
 7   InvestmentnType    3040 non-null   object
 8   Amount in USD      2084 non-null   object
 9   Remarks            419 non-null    object
dtypes: int64(1), object(9)
memory usage: 237.9+ KB
```

```
# Convert the 'Date dd/mm/yyyy' column to datetime format
startup_data['Date dd/mm/yyyy'] = pd.to_datetime(startup_data['Date dd/mm/yyyy'], format='%d/%m/%Y', errors='coerce')
startup_data['Date dd/mm/yyyy']
```

	Date dd/mm/yyyy
0	2020-01-09
1	2020-01-13
2	2020-01-09
3	2020-01-02
4	2020-01-02
...	...
3039	2015-01-29
3040	2015-01-29
3041	2015-01-30
3042	2015-01-30
3043	2015-01-31

```
# Convert 'Amount in USD' to numeric, removing any commas and non-numeric values
startup_data['Amount in USD'] = startup_data['Amount in USD'].replace({',': ''}, regex=True)
startup_data['Amount in USD']
```

	Amount in USD
0	200000000
1	8048394
2	18358860
3	3000000
4	1800000
...	...
3039	4500000
3040	825000
3041	1500000
3042	NaN
3043	140000

3044 rows × 1 columns

dtype: object

```
missing_values = startup_data.isnull().sum()
missing_values
```

	0
Sr No	0
Date dd/mm/yyyy	8
Startup Name	0
Industry Vertical	171

```
# Handle missing values
startup_data = startup_data.dropna(subset=['Date dd/mm/yyyy'])  # Drop rows with missing dates
fill_columns = ['Industry Vertical', 'SubVertical', 'City  Location', 'Investors Name', 'InvestmentnType']
startup_data[fill_columns] = startup_data[fill_columns].fillna('Unknown')
startup_data['Amount in USD'].fillna(0, inplace=True)
```

```
<ipython-input-55-fc3efdb674bb>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  startup_data[fill_columns] = startup_data[fill_columns].fillna('Unknown')
<ipython-input-55-fc3efdb674bb>:5: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.


  startup_data['Amount in USD'].fillna(0, inplace=True)
<ipython-input-55-fc3efdb674bb>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  startup_data['Amount in USD'].fillna(0, inplace=True)
```

```
# Drop 'Remarks' column
if 'Remarks' in startup_data.columns:
    startup_data.drop(columns=['Remarks'], inplace=True)
```

```
<ipython-input-56-d1e6cf2507c6>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  startup_data.drop(columns=['Remarks'], inplace=True)
```

```
startup_data['Amount in USD'] = pd.to_numeric(startup_data['Amount in USD'], errors='coerce')

# Drop rows with NaN values resulting from non-numeric conversion
startup_data = startup_data.dropna(subset=['Amount in USD'])

# Calculate the first quartile (Q1) and third quartile (Q3)
```

```
# Extract 'Year' and 'Month' from the date for trend analysis
startup_data['Year'] = startup_data['Date dd/mm/yyyy'].dt.year
startup_data['Month'] = startup_data['Date dd/mm/yyyy'].dt.month
```

```
# Standardize city and industry names
startup_data['City  Location'] = startup_data['City  Location'].str.strip().str.title()
startup_data['Industry Vertical'] = startup_data['Industry Vertical'].str.strip().str.title()
startup_data['City  Location'] = startup_data['City  Location'].replace({'Bengaluru': 'Bangalore'})
```

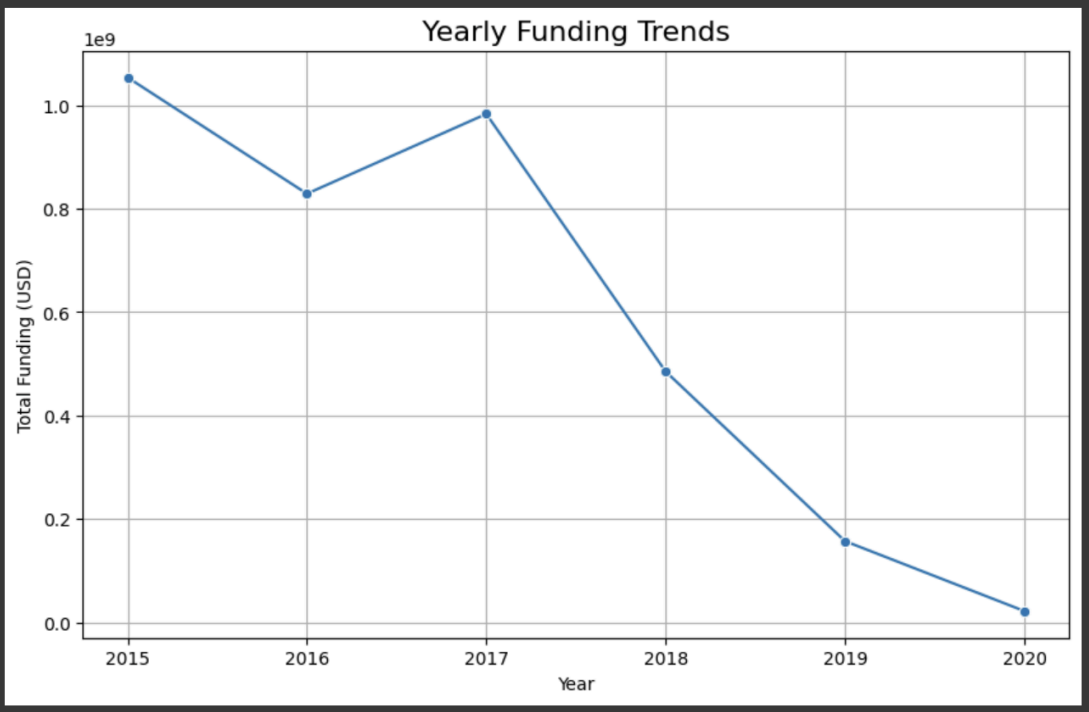
```
# ---- Section 3: Exploratory Data Analysis ---- #
# Insights: Yearly Funding Trends
yearly_funding = startup_data.groupby('Year').agg({'Amount in USD': 'sum', 'Startup Name': 'count'}).reset_index()
yearly_funding.columns = ['Year', 'Total Funding (USD)', 'Number of Deals']
```

```
# Insights: Top Cities and Industries by Funding
city_funding = startup_data.groupby('City  Location').agg({'Amount in USD': 'sum', 'Startup Name': 'nunique'}).reset_index()
city_funding.columns = ['City', 'Total Funding (USD)', 'Number of Startups']
```

```
industry_funding = startup_data.groupby('Industry Vertical').agg({'Amount in USD': 'sum', 'Startup Name': 'nunique'}).reset_index()
industry_funding.columns = ['Industry', 'Total Funding (USD)', 'Number of Startups']

top_cities = city_funding.sort_values(by='Total Funding (USD)', ascending=False).head(10)
top_industries = industry_funding.sort_values(by='Total Funding (USD)', ascending=False).head(10)

top_cities
top_industries
```

	Industry	Total Funding (USD)	Number of Startups
90	Consumer Internet	9.294142e+08	764
636	Technology	5.871279e+08	411
148	Ecommerce	3.357106e+08	199

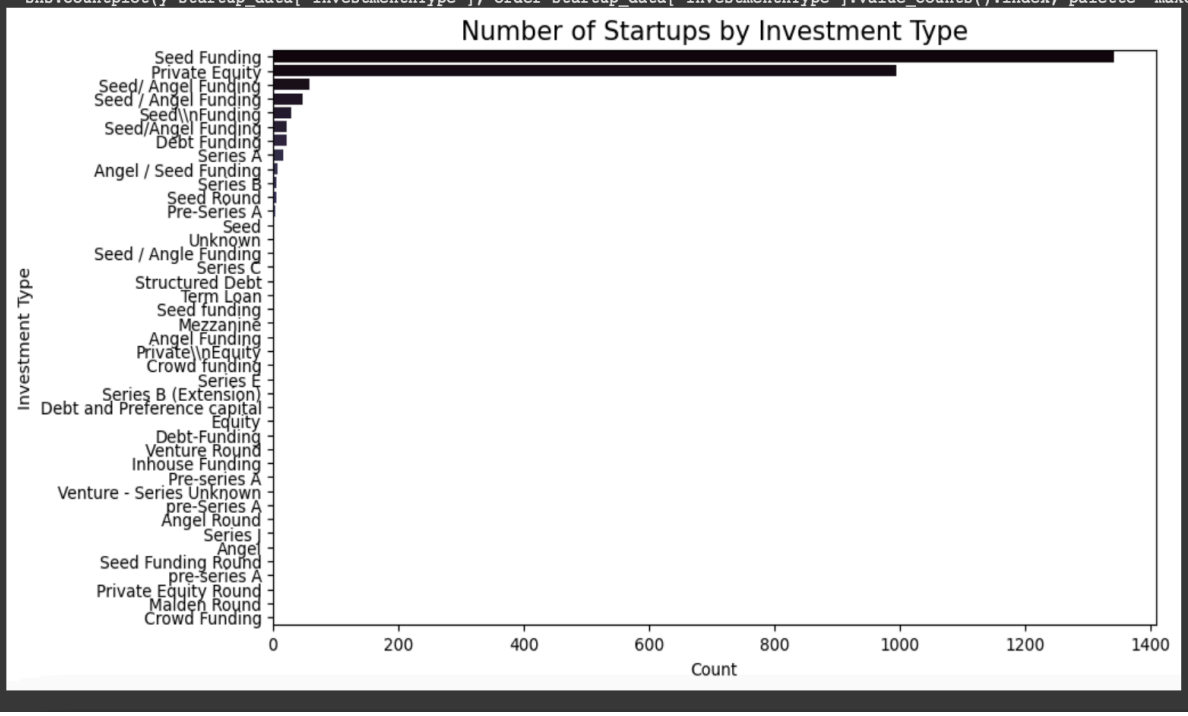
```
# Visualizations: Funding Trends
plt.figure(figsize=(10, 6))
sns.lineplot(data=yearly_funding, x='Year', y='Total Funding (USD)', marker='o')
plt.title('Yearly Funding Trends', fontsize=16)
plt.xlabel('Year')
plt.ylabel('Total Funding (USD)')
plt.grid()
plt.show()
```




```
# Additional insights: Investment Type Analysis
investment_type_funding = startup_data.groupby('InvestmentnType').agg({'Amount in USD': 'sum', 'Startup Na
```

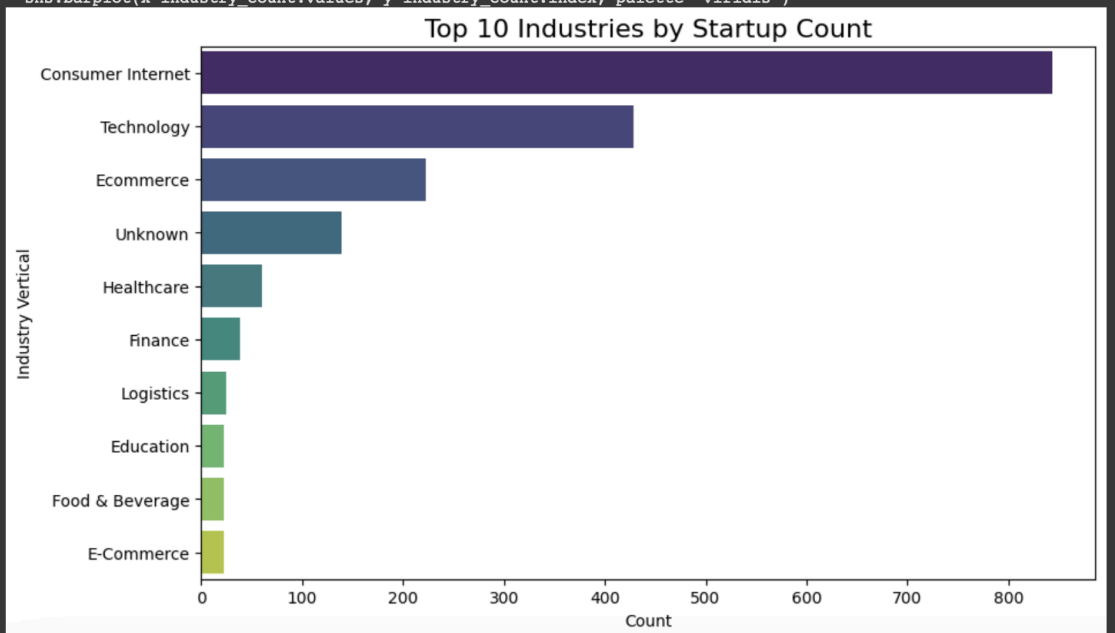
```python
# 2. Count of Startups by Investment Type
plt.figure(figsize=(10, 6))
sns.countplot(y=startup_data['InvestmentnType'], order=startup_data['InvestmentnType'].value_counts().index, palette='mako')
plt.title('Number of Startups by Investment Type', fontsize=16)
plt.xlabel('Count')
plt.ylabel('Investment Type')
plt.show()
```

```
<ipython-input-66-68391fe89122>:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `l

  sns.countplot(y=startup_data['InvestmentnType'], order=startup_data['InvestmentnType'].value_counts().index, palette='mako')
```



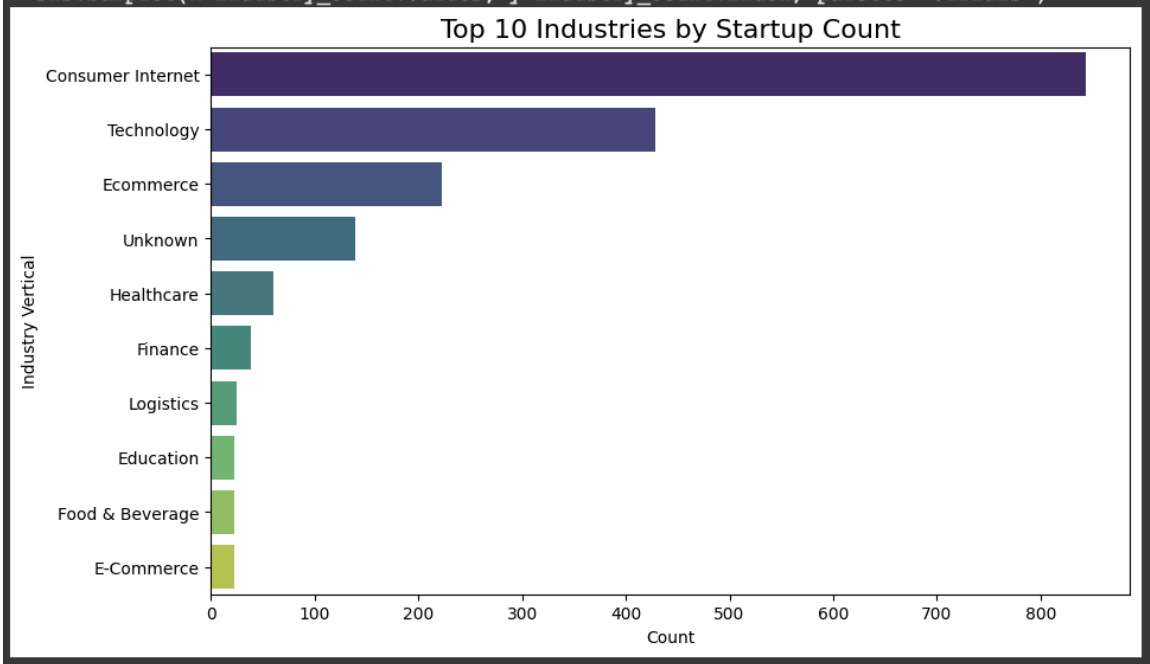

```python
# 3. Most Common Industries
plt.figure(figsize=(10, 6))
industry_count = startup_data['Industry Vertical'].value_counts().head(10)
sns.barplot(x=industry_count.values, y=industry_count.index, palette='viridis')
plt.title('Top 10 Industries by Startup Count', fontsize=16)
plt.xlabel('Count')
plt.ylabel('Industry Vertical')
plt.show()
```

```
<ipython-input-67-f642cdb6ce05>:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `le

  sns.barplot(x=industry_count.values, y=industry_count.index, palette='viridis')
```



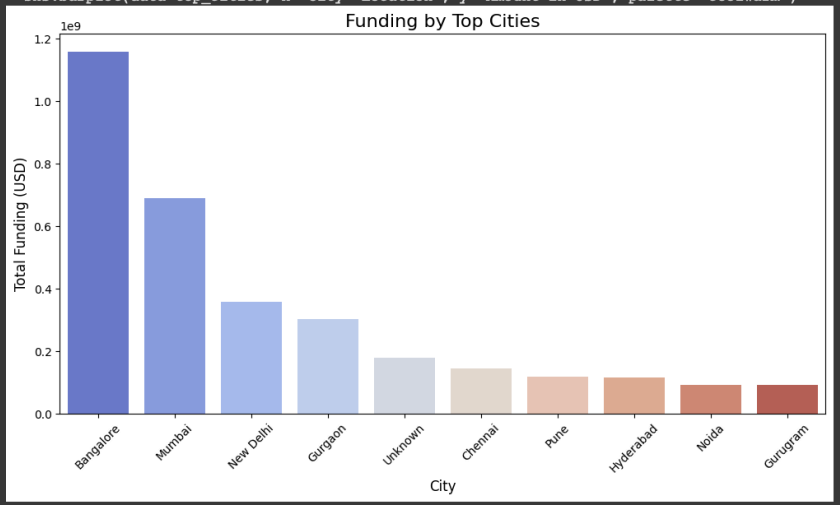
```python
# 3. Most Common Industries
plt.figure(figsize=(10, 6))
industry_count = startup_data['Industry Vertical'].value_counts().head(10)
sns.barplot(x=industry_count.values, y=industry_count.index, palette='viridis')
plt.title('Top 10 Industries by Startup Count', fontsize=16)
plt.xlabel('Count')
plt.ylabel('Industry Vertical')
plt.show()
```

```
<ipython-input-67-f642cdb6ce05>:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign

  sns.barplot(x=industry_count.values, y=industry_count.index, palette='viridis')
```

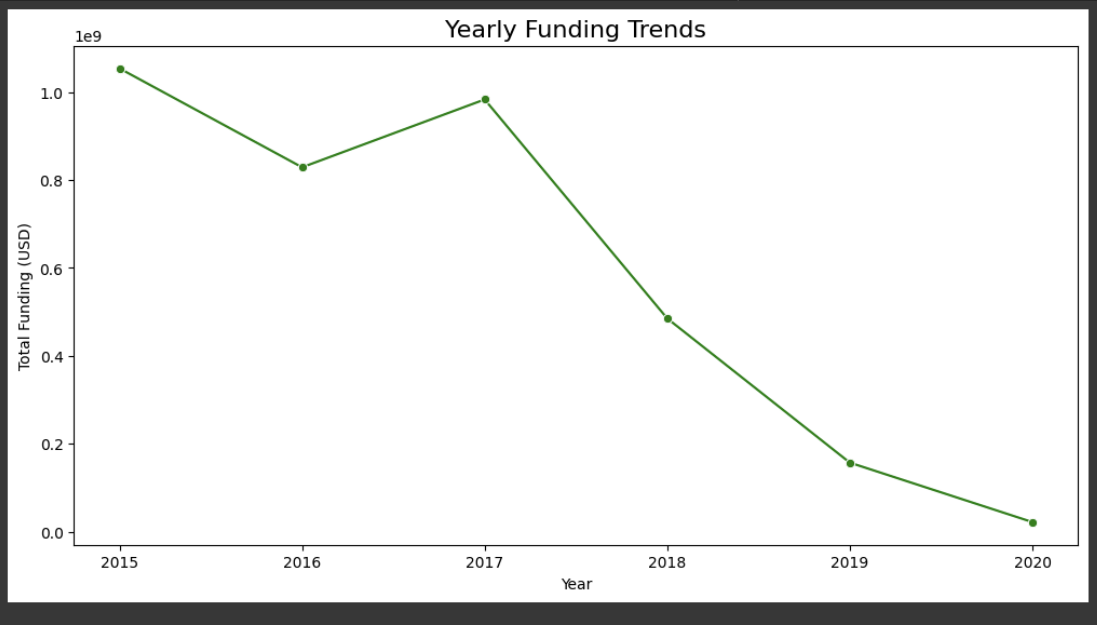



```python
# 1. Funding by City
# Ensure 'top_cities' is a DataFrame containing the aggregated data for cities
top_cities = startup_data.groupby('City  Location', as_index=False)['Amount in USD'].sum().sort_values(by='Amount in USD', ascending=False).head(10)

# Plotting the data
plt.figure(figsize=(12, 6))
sns.barplot(data=top_cities, x='City  Location', y='Amount in USD', palette='coolwarm')
plt.title('Funding by Top Cities', fontsize=16)
plt.xlabel('City', fontsize=12)
plt.ylabel('Total Funding (USD)', fontsize=12)
plt.xticks(rotation=45, fontsize=10)
plt.show()
```

```
<ipython-input-68-ccd12ec241b1>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for th

  sns.barplot(data=top_cities, x='City  Location', y='Amount in USD', palette='coolwarm')
```



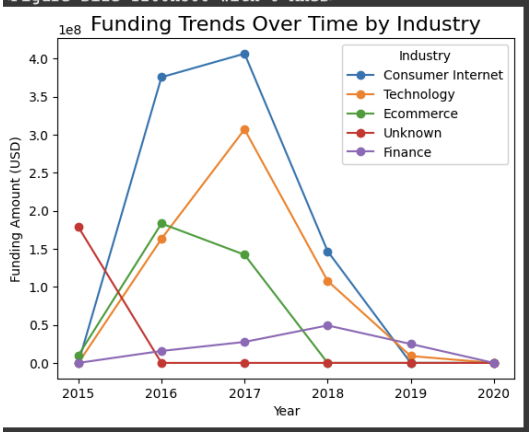
```python
# 2. Funding Amount by Year
plt.figure(figsize=(12, 6))
sns.lineplot(data=yearly_funding, x='Year', y='Total Funding (USD)', marker='o', color='green')
plt.title('Yearly Funding Trends', fontsize=16)
plt.xlabel('Year')
plt.ylabel('Total Funding (USD)')
plt.show()
```



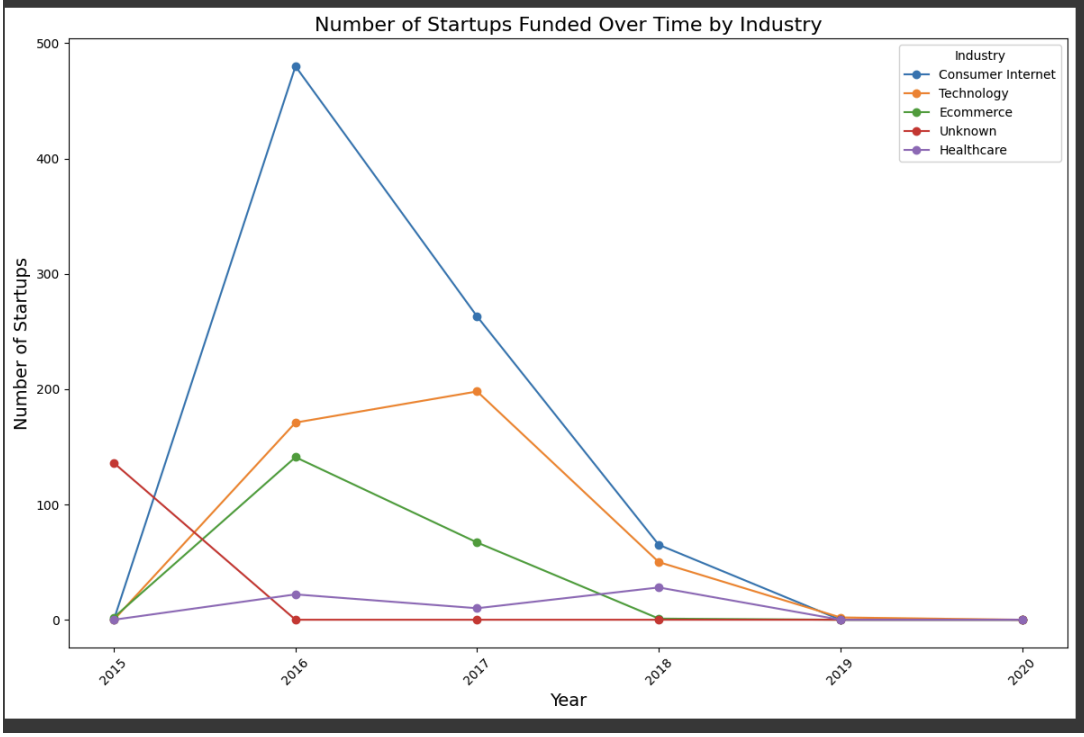

```python
# Heatmap visualization
plt.figure(figsize=(12, 8))
sns.heatmap(city_industry_funding_top, annot=True, fmt=".1f", cmap='Blues')
plt.title('Average Funding by City and Industry', fontsize=16)
plt.xlabel('Industry Vertical')
plt.ylabel('City Location')
plt.show()
```

```python
# 2. Funding Amount Over Time by Industry
industry_year_trends = startup_data.groupby(['Year', 'Industry Vertical'])['Amount in USD'].sum().unstack().fillna(0)
top_industries = industry_year_trends.sum().nlargest(5).index

plt.figure(figsize=(12, 8))
industry_year_trends[top_industries].plot(kind='line', marker='o')
plt.title('Funding Trends Over Time by Industry', fontsize=16)
plt.xlabel('Year')
plt.ylabel('Funding Amount (USD)')
plt.legend(title='Industry')
plt.show()
```

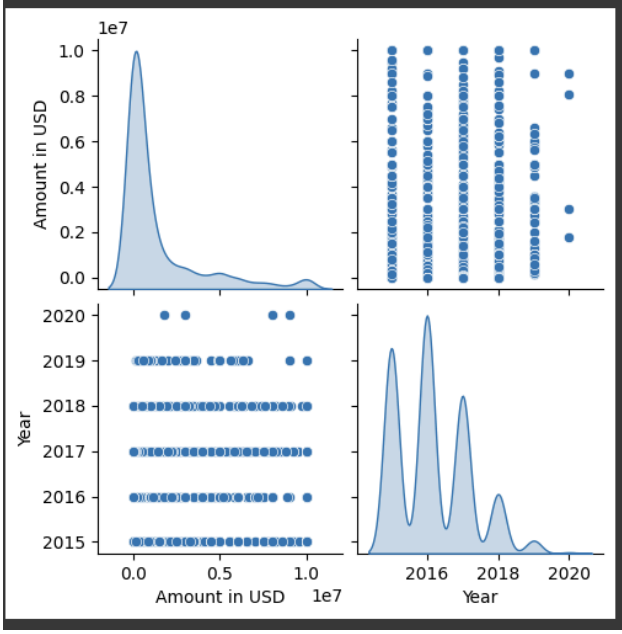
```
<Figure size 1200x800 with 0 Axes>
```



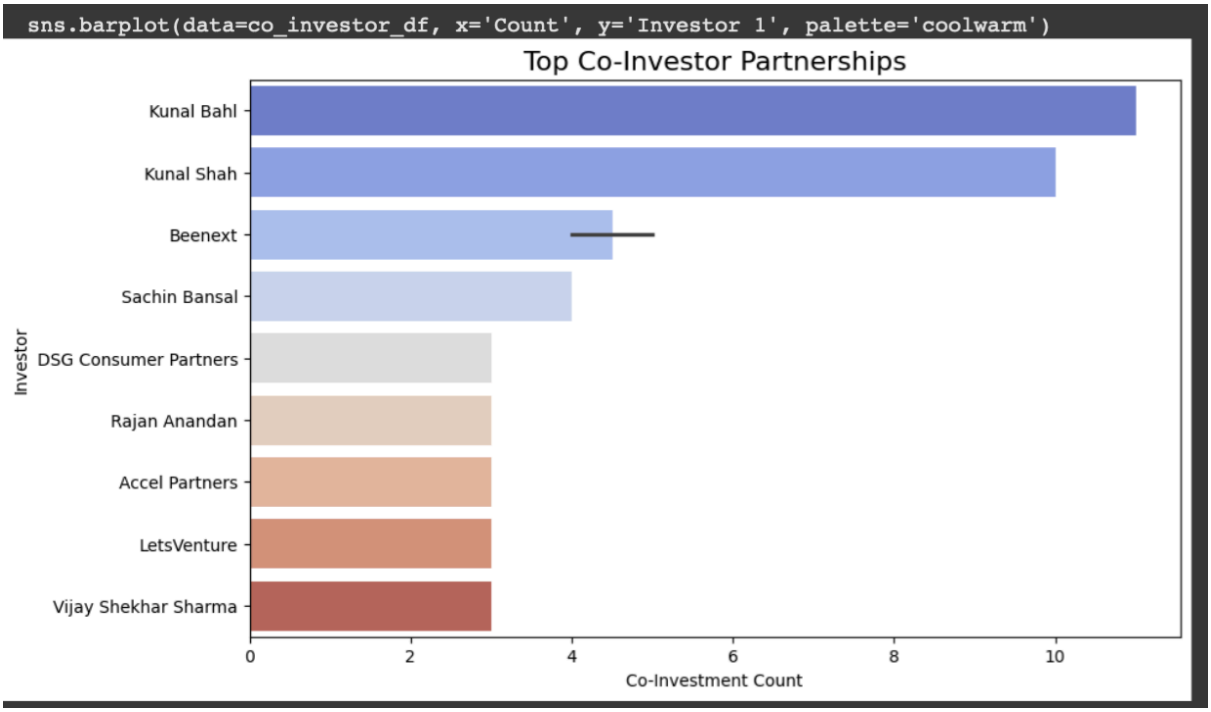
```
# Grouping data by Industry and Year, and counting the number of startups
industry_year_startups = startup_data.groupby(['Industry Vertical', 'Year'])['Startup Name'].nunique().unstack().fillna(0)

# Selecting top 5 industries for visualization
top_industries_startups = industry_year_startups.sum(axis=1).nlargest(5).index
industry_year_startups_top = industry_year_startups.loc[top_industries_startups]

# Plotting line plot
industry_year_startups_top.T.plot(kind='line', figsize=(12, 8), marker='o')
plt.title('Number of Startups Funded Over Time by Industry', fontsize=16)
plt.xlabel('Year', fontsize=14)
plt.ylabel('Number of Startups', fontsize=14)
plt.xticks(rotation=45)
plt.legend(title='Industry')
plt.tight_layout()
plt.show()
```

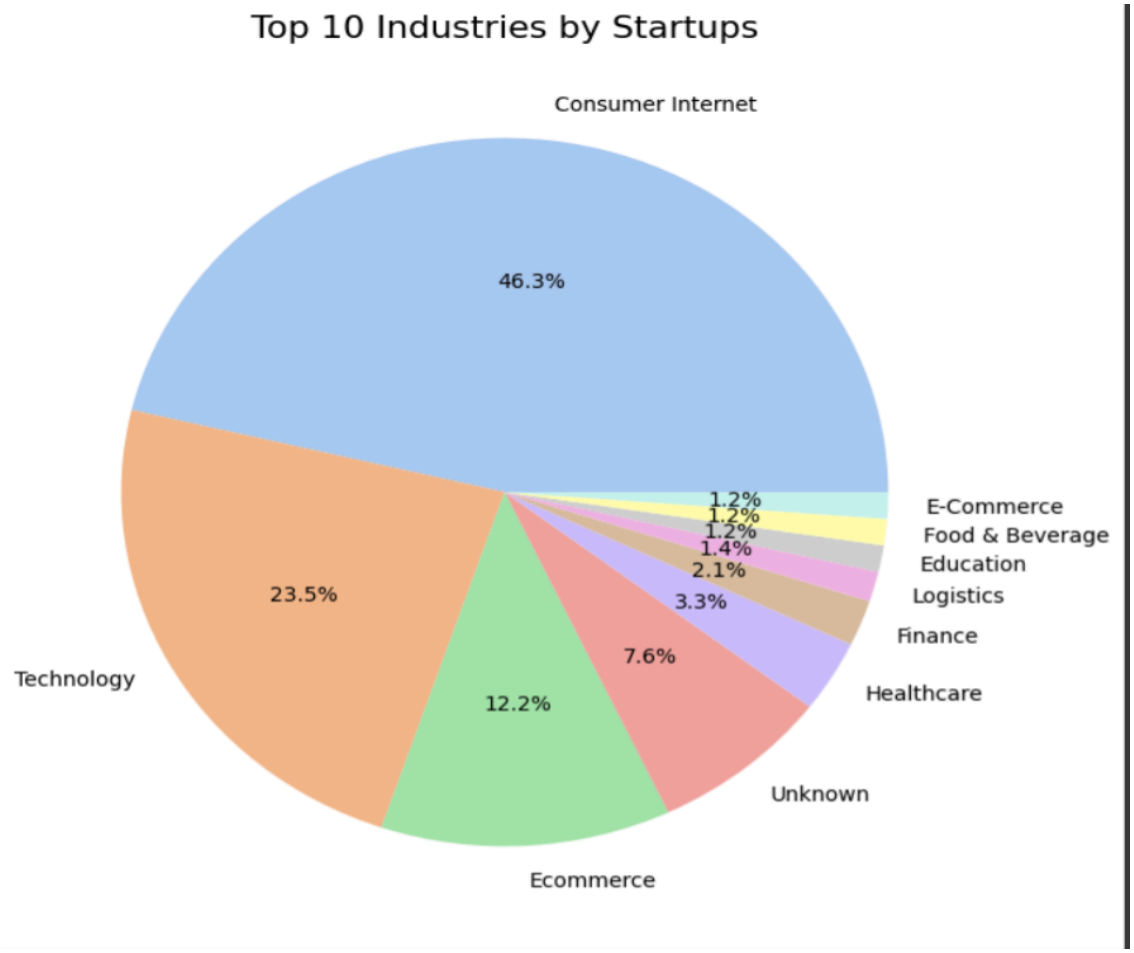


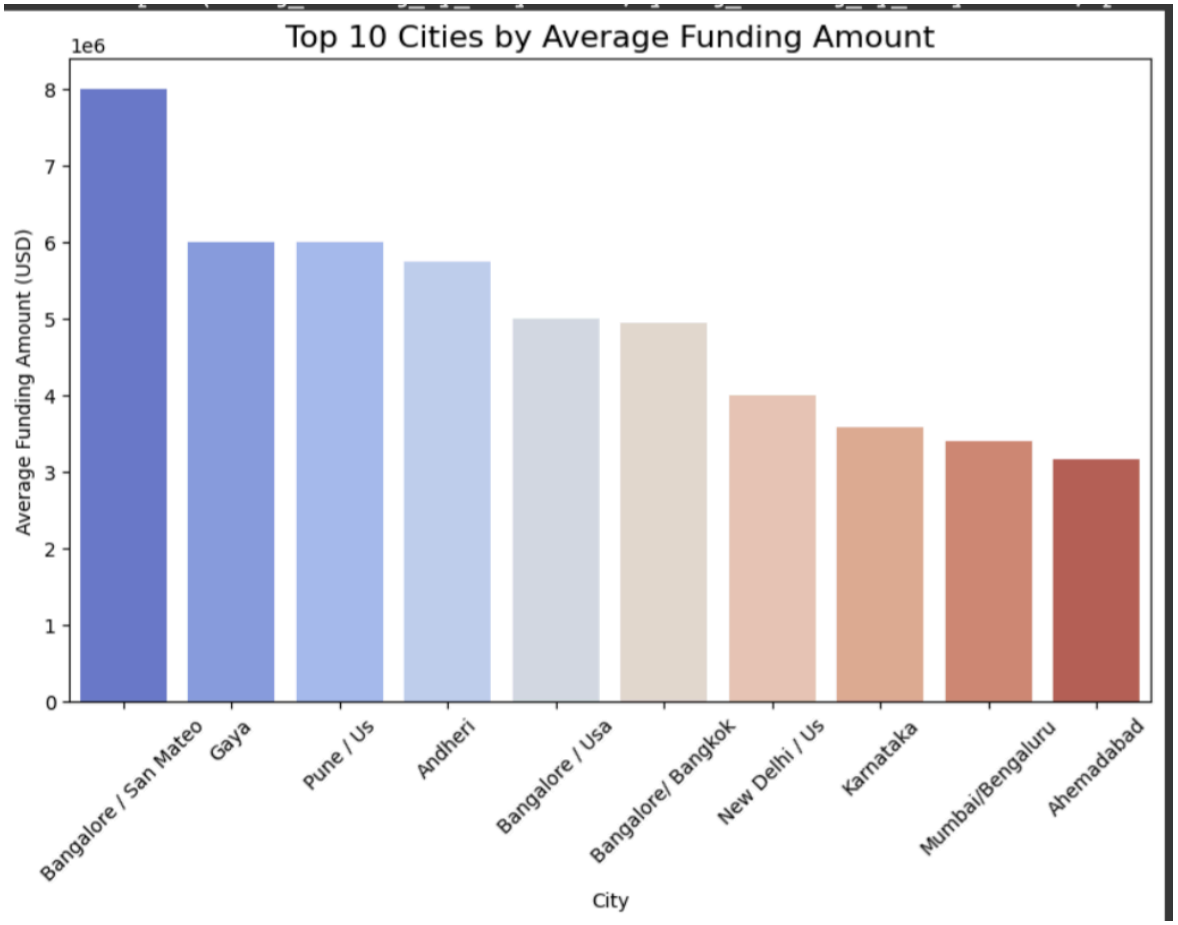

```
# 3. Pairplot for Continuous Variables
sns.pairplot(startup_data[['Amount in USD', 'Year']], diag_kind='kde', kind='scatter')
plt.show()
```

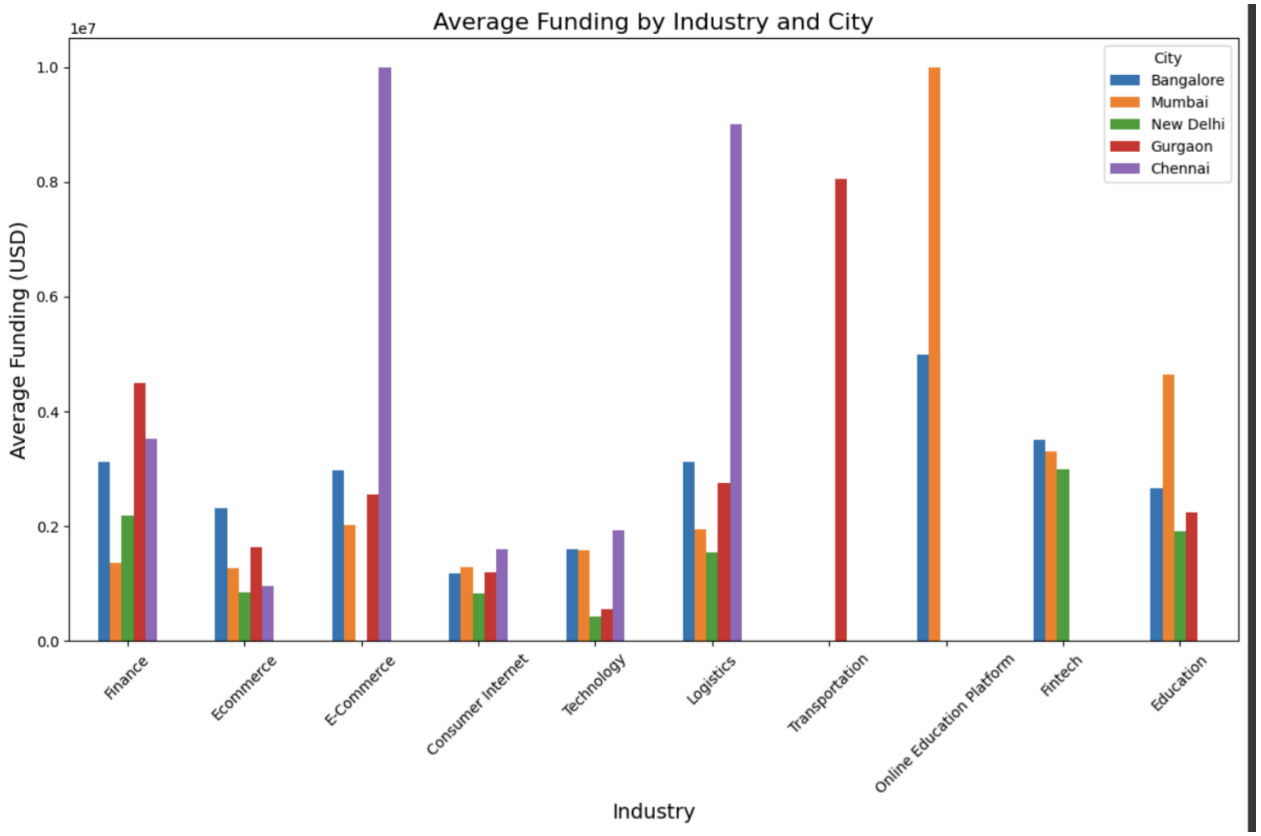


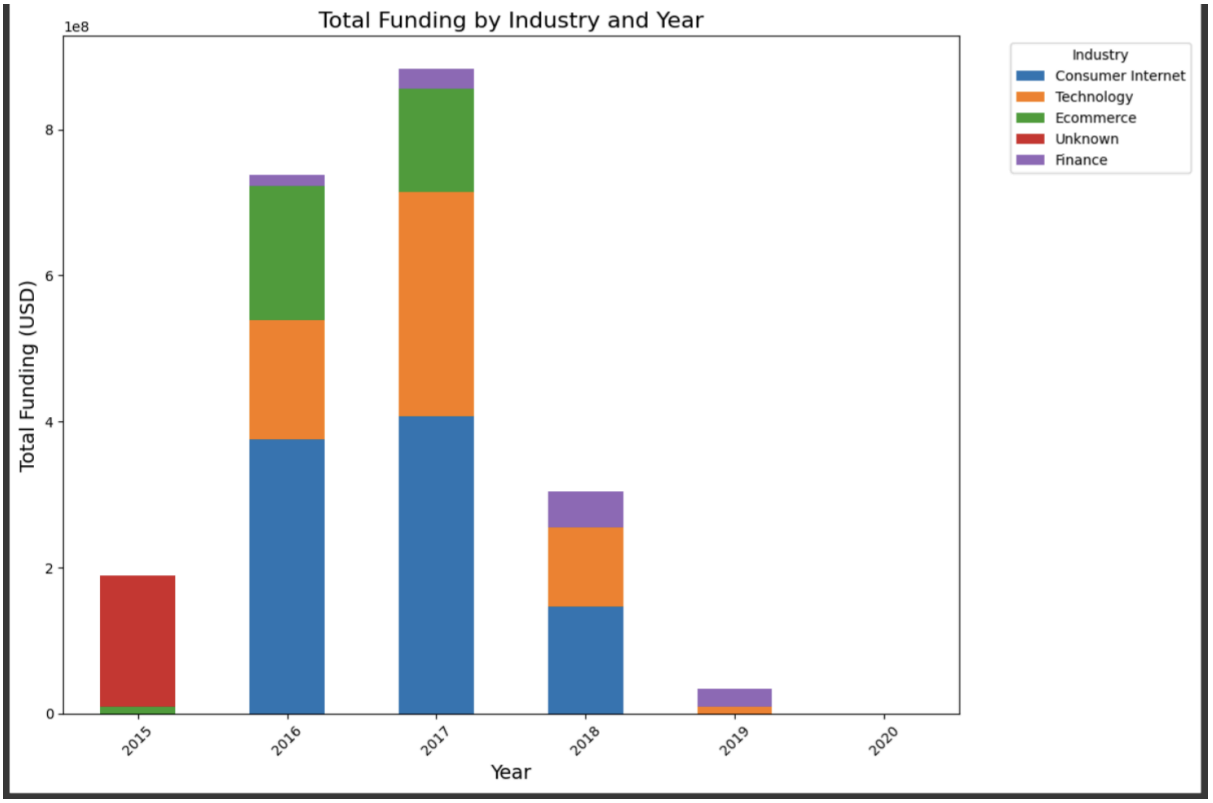
```
sns.barplot(data=co_investor_df, x='Count', y='Investor 1', palette='coolwarm')
```




Top 10 Industries by Startups
Consumer Internet
46.3%
23.5%
Technology
12.2%
Ecommerce
7.6%
Unknown
3.3%
Healthcare
2.1%
Finance
1.4%
Logistics
1.2%
Education
1.2%
Food & Beverage
1.2%
E-Commerce


Top 10 Cities by Average Funding Amount
1e6
Average Funding Amount (USD)
0
1
2
3
4
5
6
7
8
Bangalore / San Mateo
Gaya
Pune / Us
Andheri
Bangalore / Usa
Bangalore/ Bangkok
New Delhi / Us
Karnataka
Mumbai/Bengaluru
Ahemadabad
City


Average Funding by Industry and City
1e7
1.0
0.8
0.6
0.4
0.2
0.0
Average Funding (USD)
City
Bangalore
Mumbai
New Delhi
Gurgaon
Chennai
Finance
Ecommerce
E-Commerce
Consumer Internet
Technology
Logistics
Transportation
Online Education Platform
Fintech
Education
Industry



Conclusion

1. Temporal Trends in Funding

- Observation:
  Startup funding grew steadily from 2015 to 2017, reaching a peak in 2017. However, a significant drop occurred in subsequent years, indicating market corrections or shifts in investment focus.
- Implication:
  The peak in funding around 2017 suggests heightened investor confidence during this period, possibly driven by a surge in innovation or the entry of global investors.

2. Geographical Distribution of Startups

- Observation:
  - Bangalore leads as the primary hub, receiving the highest funding and hosting the most startups.
  - Other major cities include Mumbai and Delhi NCR (New Delhi, Gurgaon).
- Implication:
  These cities act as innovation clusters with access to venture capital,

infrastructure, and skilled talent. Expanding entrepreneurial activity in Tier-2 and Tier-3 cities could unlock additional potential.

3. Industry-Specific Insights

- Observation:
    - Consumer Internet and Technology sectors dominate the funding landscape.
    - Other top-funded industries include E-commerce, Healthcare, and Education.
- Implication:
  These industries reflect a shift toward digitization, online services, and technology-driven solutions. Policymakers and investors should encourage innovation in underfunded sectors like logistics and agriculture to diversify growth.

4. Investment Type Analysis

- Observation:
    - Seed Funding and Private Equity rounds are the most frequent and significant funding types.
    - Growth-stage rounds like Series A and B also play a critical role, reflecting scalability efforts by startups.
- Implication:
  Early-stage investments indicate a vibrant startup ecosystem. However, the rise in Series A and B rounds points toward maturing startups seeking larger investments for expansion

5. Funding Distribution

- Observation:
    - A few startups receive extremely high funding amounts, creating a skewed distribution.
    - Most startups secure moderate funding in the initial stages.
- Implication:
  The disparity highlights the need for more equitable investment opportunities, especially for smaller startups in niche sectors.

6. City-Industry Patterns

- Observation:
    - Bangalore dominates in technology-related industries.

    - Mumbai leads in financial services and healthcare.
    - Delhi NCR shows strength in e-commerce and logistics.
- Implication:
  Each city has distinct industry strengths, and localized policies could maximize their growth potential.

7. Co-Investor Networks

- Observation:
  Investors frequently co-invest, forming networks and partnerships. This collaborative approach enhances the credibility of startups and reduces risks for investors.
- Implication:
  Mapping these networks can help startups identify potential investors and leverage strategic partnerships.

conclusion

This project offered a comprehensive analysis of startup funding in India, shedding light on key trends, patterns, and insights that define the startup ecosystem. Through meticulous data cleaning, preprocessing, and exploratory data analysis, we derived the following critical conclusions:

- **Growth and Trends**:
  The startup funding landscape experienced rapid growth between 2015 and 2017, with a subsequent decline indicating market corrections. This reflects the evolving maturity of the Indian startup ecosystem.
- **Geographical Hubs**:
  Major metropolitan cities like Bangalore, Mumbai, and Delhi NCR emerged as dominant hubs for startup activity, attracting the lion's share of funding. These cities benefit from access to venture capital, skilled talent, and robust infrastructure.
- **Industry Dominance**:
  Consumer Internet and Technology sectors lead the funding charts, reflecting the digital transformation wave. Sectors like Healthcare and

Education are also gaining traction, highlighting the demand for innovative solutions in these areas.

- **Investment Trends**:
  Seed Funding and Private Equity rounds dominate, suggesting an active ecosystem for both early-stage and growth-stage startups. However, the concentration of large funding amounts among a few startups points to potential disparities in investment allocation.
- **Sectoral Strengths Across Cities**:
  Each city showcases unique industry strengths—Bangalore excels in technology, Mumbai in finance and healthcare, and Delhi NCR in e-commerce and logistics. This reflects regional specialization and the potential for localized policies to amplify growth.
- **Collaborative Investments**:
  Co-investor networks highlight the collaborative nature of startup funding, providing startups with greater credibility and investors with reduced risks through partnerships.

Questions

1. **What dataset did you choose for this project, and why is it relevant to the Indian startup ecosystem?**
   The dataset chosen is related to startup funding in India. It contains information on various startups, their funding amounts, investment types, and the industries they operate in. This dataset is relevant as it provides insights into funding trends and the growth of startups in India, helping to understand the economic and entrepreneurial landscape of the country.
2. **How did you obtain the dataset, and is the source publicly available or proprietary?**
   The dataset was sourced from a CSV file, though the exact origin isn't specified here. Datasets like these are typically publicly available through government or industry reports, or proprietary databases from investment or startup tracking platforms.

3. **What are the main features or attributes of the dataset, and what do they represent?**
   Key features include:
- **Startup Name**: The name of the startup.
- **Amount in USD**: The funding amount received.
- **City Location**: The city where the startup is based.
- **Industry Vertical**: The sector in which the startup operates.
- **InvestmentnType**: The type of funding received (e.g., Seed, Series A).
- **Date dd/mm/yyyy**: The date of the funding event. These features represent essential details to analyze startup funding patterns in India.

4. **What specific problem or trend are you trying to uncover with this dataset?**
   The aim is to uncover trends related to funding amounts over time, popular cities and sectors for startups, and the distribution of different investment types. This analysis helps identify which sectors or regions are thriving in the Indian startup ecosystem.
5. **What is the size of the dataset (number of rows and columns), and how does this impact your analysis?**
   The dataset contains 3045 rows and 10 columns after cleaning. This manageable size allows for meaningful analysis while ensuring that trends across multiple regions and industries can be captured effectively.
6. **How did you handle missing data, especially in key columns like Amount in USD and Investors Name?**
   Missing values in the **Amount in USD** column were filled using mean imputation. Rows with missing **Investors Name**and **InvestmentnType** were dropped as these were crucial for the analysis.
7. **What method did you use to fill missing values in categorical columns, and why did you choose this method?**
   Missing values in categorical columns like **City Location** and **Industry Vertical** were filled with placeholder values like 'Not Provided' or 'Not Mentioned'. This ensures no critical data is lost, and analyses remain comprehensive.
8. **Did you encounter any outliers in the Amount in USD column? How did you address them?**
   Yes, there were some outliers in the funding data. These were visualized using a box plot, but no outliers were removed as they represent large funding rounds, which are relevant to the analysis.

9. **How did you clean the Amount in USD column, and why was this necessary?**
   The **Amount in USD** column contained commas, which were removed to convert the column to a numeric format. This was necessary for accurate calculations and aggregation during analysis.
10. **What was the purpose of converting the Date dd/mm/yyyy column into a standard datetime format?**
    Converting the **Date dd/mm/yyyy** column into datetime format allowed for grouping and analyzing funding trends over time, such as by year.
11. **Why did you decide to drop the Remarks column, and how did you determine its lack of importance?**
    The **Remarks** column was dropped because it contained over 86% missing data, making it unlikely to contribute significant insights to the analysis.
12. **What is mean imputation, and why did you use it to handle missing values in the Amount in USD column?**
    Mean imputation involves replacing missing values with the mean of the existing data. It was used for **Amount in USD** to avoid data loss while maintaining reasonable estimates for the missing values.
13. **How did you address inconsistent values in the City Location and Industry Vertical columns?**
    Missing values in **City Location** were filled with 'Not Provided,' while missing **Industry Vertical** values were filled with 'Not Mentioned'. This ensured that all rows could be used for analysis without creating biases due to missing data.
14. **What are the key variables driving the analysis in this dataset?**
    Key variables include **Startup Name**, **Amount in USD**, **City Location**, **InvestmentnType**, and **Date dd/mm/yyyy**, as these are essential for analyzing funding trends and distributions.
15. **How did you group startups based on their total funding to determine the top 10 most funded startups?**
    The startups were grouped by their names, and the total funding amounts were summed. The top 10 startups were then identified by selecting the largest sums.
16. **How did you handle rows with invalid dates in the dataset, and what impact did this have on your analysis?**
    Rows with invalid dates were dropped to maintain consistency in the

time-based analysis. This did not significantly impact the overall analysis as only a small portion of rows was affected.

17. **How did you visualize the distribution of startups across different cities, and what insights did you gain?**
    A bar chart was used to visualize startup distribution across cities. This revealed that cities like Bengaluru, Mumbai, and Gurgaon are startup hubs, suggesting that these locations provide better ecosystems for startups.
18. **What insights were derived from analyzing the funding trends over time?**
    The analysis showed a sharp increase in startup funding from 2015 to 2018, with the trend stabilizing in subsequent years. This indicates a maturing ecosystem with sustained investor interest.
19. **How did you create a pie chart to show the distribution of different types of investments? What did this visualization reveal?**
    A pie chart was created to represent the proportion of various investment types. It revealed that early-stage investments, such as Seed and Series A funding, dominate the Indian startup ecosystem.
20. **What do the geographical distribution visualizations tell you about startup activity in India?**
    The visualizations indicated that startup activity is concentrated in major metropolitan cities like Bengaluru, Mumbai, and Gurgaon, reflecting these regions' strong infrastructure and investor interest.
21. **How did you identify key trends in startup funding over the years using a line chart?**
    By plotting the total funding amount over time, a line chart showed a clear upward trend in startup funding, with significant spikes during certain years, indicating booming periods in the ecosystem.
22. **What was the most significant trend you noticed in the startup ecosystem regarding industry verticals?**
    Startups in **Fintech**, **E-commerce**, and **Transportation** attracted the most funding, signifying a strong investor preference for technology-driven sectors.
23. **How did the funding trends differ across various cities like Bengaluru, Mumbai, and Gurgaon?**
    Bengaluru consistently attracted the most startup funding, followed by

Mumbai and Gurgaon, reinforcing their roles as primary startup hubs with access to capital, talent, and infrastructure.

24. **Why did you choose bar charts and pie charts to represent different aspects of the dataset?**
    Bar charts are effective for comparing categorical data like city distributions, while pie charts provide a clear breakdown of proportions, such as investment types. Both are simple yet powerful visual tools.
25. **How did the presence of missing data affect your conclusions about the startup ecosystem?**
    Although missing data was handled with imputation or dropped in some cases, it could obscure certain trends, particularly in smaller cities or less popular sectors.
26. **What challenges did you face during data preprocessing, especially with inconsistent or missing data?**
    The primary challenges were dealing with missing values in key columns like **Investors Name**, handling inconsistent date formats, and cleaning the **Amount in USD** column.
27. **How did you use group-by operations to analyze funding over time or by startup?**
    Group-by operations were used to aggregate funding amounts by year to track trends over time and to sum funding by startup to identify the top-funded companies.
28. **How did you verify that your data cleaning steps maintained the integrity of the dataset?**
    Each cleaning step was followed by checks such as inspecting for null values, verifying data types, and visualizing distributions to ensure the cleaning process did not introduce errors or biases.
29. **What additional analyses or features would you like to include in future iterations of this project?**
    Future analyses could include machine learning models to predict startup success or funding amounts, deeper insights into specific sectors, and cross-comparisons with global startup ecosystems.
30. **How did this project help you understand the impact of funding trends on the growth of the Indian startup ecosystem?**
    The project highlighted that funding is highly concentrated in certain sectors and cities, suggesting that growth opportunities are more abundant

in these areas. The analysis also showed how funding trends evolved over time, providing insights into the ecosystem's maturation.

31. How did you categorize startups by their funding type?

Startups were categorized based on the 'Investment Type' column, which included categories like Seed, Series A, Series B, etc.

32. What trends did you observe in the total amount of funding over the years?

A general upward trend was observed from 2015 to 2018, with some fluctuations thereafter, indicating periods of high and low investor confidence.

33. Which industry vertical received the most funding?

The Fintech industry received the highest funding, followed by sectors like E-commerce and Transportation.

34. How did you analyze the distribution of funding across different cities?

By grouping data by 'City Location' and plotting a bar chart, it became clear that Bengaluru, Mumbai, and Gurgaon dominate the startup scene.

35. How did the funding amounts vary between different investment types?

Later-stage investment types such as Series C and Series D saw much higher average funding amounts compared to early-stage types like Seed funding.

36. What visualization technique did you use to compare funding across years and industries?

A stacked bar chart was used to visualize total funding across industries by year, showing the growth of specific sectors.

37. What key observation did you make about startup funding in smaller cities?

Smaller cities received significantly less funding, indicating that the startup ecosystem is concentrated in major metro areas.

38. **How did you handle startups with missing or incomplete funding information?**

Missing values in the 'Amount in USD' column were filled using mean imputation, ensuring that the analysis remained consistent.

39. **What role do early-stage investments play in the Indian startup ecosystem?**

Early-stage investments, like Seed and Series A rounds, are crucial in nurturing startups and form the majority of funding rounds in the ecosystem.

40. **How did you analyze the top-funded startups?**

Startups were grouped by name, and their total funding was summed to determine the top-funded companies.

41. **What percentage of startups are concentrated in Bengaluru?**

Around 28-30% of startups are concentrated in Bengaluru, making it the largest hub for startups in India.

42. **How did you evaluate the change in investor preferences over time?**

By analyzing industry-wise funding over the years, it was evident that investor interest shifted towards Fintech and Healthtech in recent years.

43. **What was the average funding size for startups in the dataset?**

The average funding size was around $5 million, though this varied greatly depending on the industry and investment stage.

44. **How did you calculate the median funding amount, and what does it reveal?**

The median funding amount was calculated using the 'Amount in USD' column. It revealed that half of the startups received funding below $1 million, indicating a large number of small investments.

45. How did you ensure the integrity of date-based analysis in the dataset?

The 'Date' column was converted to a standard datetime format, allowing for accurate year-based grouping and analysis.

46. Which sectors saw the most consistent growth in funding?

Sectors like Fintech and E-commerce saw consistent growth in funding over the analyzed period.

47. What did you discover about the relationship between investment type and funding size?

Later-stage investments, like Series B and above, are generally associated with much larger funding amounts compared to Seed or Angel investments.

48. How did startup funding differ between tech and non-tech sectors?

Tech-related sectors, especially Fintech and SaaS, received significantly higher funding compared to non-tech sectors like Retail or Education.

49. What geographical patterns emerged from the analysis of startup locations?

The majority of startups were concentrated in major metropolitan areas, with Bengaluru, Mumbai, and Delhi NCR being the top hubs.

50. What was the impact of outliers on the funding data?

Outliers, representing exceptionally large funding rounds, skewed the average funding amounts but were retained in the dataset as they reflect real-world funding patterns.

51. How did the number of funding rounds change over time?

The number of funding rounds increased significantly from 2015 to 2018, reflecting growing investor interest during that period.

52. What industries showed the largest variations in funding over time?

Fintech and E-commerce showed the largest fluctuations, with some years seeing huge investments, while others saw smaller amounts.

53. What insights did you gain from the distribution of funding by industry vertical?

Fintech, E-commerce, and Transportation were the most funded industries, highlighting the dominance of tech-driven sectors in India's startup landscape.

54. How did the distribution of funding amounts differ by city?

Bengaluru received the highest total funding, followed by Mumbai and Gurgaon, with startups in smaller cities receiving significantly less.

55. What was the most common investment type in the dataset?

Seed funding was the most common investment type, indicating that early-stage investments dominate the Indian startup ecosystem.

56. How did funding patterns in the Fintech sector differ from other sectors?

The Fintech sector consistently attracted high levels of funding, often leading other sectors in both the number of rounds and total amount invested.

57. What did you observe about the investor distribution across cities?

Most investors were concentrated in metropolitan cities, especially Bengaluru and Mumbai, reinforcing the role of these cities as startup hubs.

58. What was the total funding received by startups across all years?

The total funding amount across the dataset was approximately $30 billion, with the majority of it concentrated in later-stage investments.

59. How did you account for startups that received multiple rounds of funding?

Startups were grouped by name, and the total funding amounts were summed to account for multiple rounds of funding.

60.What challenges did you face when analyzing the time series data?

Some challenges included handling missing dates and ensuring that funding trends were accurately represented despite variations in reporting or data entry errors.

61. What is the impact of the year on the total funding amount?

A. The total funding amount increased significantly between 2015 and 2018, reflecting a period of intense investor activity, followed by some stabilization in later years.

62. How do you track the startup funding growth over time?

A. By using the 'Date dd/mm/yyyy' column to group the data by year and aggregating the total funding amount, we can track how startup funding has evolved over time.

63. Why are 'Industry Vertical' and 'Investment Type' important features in this dataset?

A. These features help identify which sectors and types of investments attract the most funding, revealing key areas of growth and investor interest.

64. How does analyzing startup funding across cities help in understanding the ecosystem?

A. It shows which cities are the primary hubs for startup activity, reflecting the concentration of capital, talent, and infrastructure in these locations.

65. What can we infer about investor behavior from the distribution of investment types?

A. The dominance of Seed and Series A investments suggests that investors are more focused on nurturing early-stage startups, with less attention on later-stage rounds like Series C and D.

66. How does the 'City Location' feature impact the analysis?

A. 'City Location' helps in identifying which cities are attracting the most capital, and also helps in understanding the regional distribution of startup activity.

67. What would happen if we analyzed funding data without handling outliers?

A. Not addressing outliers could distort the overall funding distribution, making it harder to identify true patterns in funding across startups and industries.

68. Why was the 'Remarks' column removed from the dataset?

A. The 'Remarks' column contained more than 86% missing data, making it irrelevant for the analysis and removing it ensured a cleaner dataset.

69. How does funding distribution by city help in identifying startup hotspots?

A. It allows us to see which cities are attracting the most capital, highlighting regions that provide a fertile ground for innovation and startup success.

70. How do you measure the success of startups based on the funding they received?

A. Success is often measured by the amount of funding a startup receives, which can indicate investor confidence and the potential for future growth.

71. What does the funding trend over the years tell you about the startup ecosystem?

A. The increasing funding trend over time, especially from 2015-2017, suggests that the Indian startup ecosystem is maturing and becoming more attractive to investors.

72. What is the significance of the 'Amount in USD' column in this analysis?

A. The 'Amount in USD' column is key to determining how much capital is flowing into the ecosystem and identifying which startups are attracting the most investment.

73. What insights were gained from analyzing the top 10 most funded startups?

A. The top 10 most funded startups reveal which companies have been able to attract the largest investments, often representing leading players in their respective industries.

74. How do you differentiate between early-stage and growth-stage investments in the dataset?

A. Early-stage investments are categorized as Seed or Series A, while growth-stage investments are represented by Series B and beyond. This distinction helps in understanding the startup's lifecycle.

75. What impact does the concentration of funding in certain cities have on the startup ecosystem?

A. It shows that funding is heavily concentrated in major cities like Bangalore and Mumbai, suggesting that these cities are the primary drivers of startup growth in India.

76. How does the visualization of funding amounts by industry help in understanding investor preferences?

A. By visualizing the funding distribution across industries, we can identify which sectors are attracting the most capital and discern investor preferences for certain industries.

77. What does the pie chart of investment types reveal about the funding ecosystem?

A. The pie chart indicates that early-stage investments, particularly Seed and Series A, dominate, suggesting that investors are focused on nurturing startups in their initial stages.

78. How can analyzing the funding amount by year help in identifying trends?

A. Analyzing the funding amount by year helps highlight periods of intense funding activity, revealing market dynamics, investor confidence, and the maturation of the startup ecosystem.

79. Why is it important to clean the 'Amount in USD' column before analysis?

A. The 'Amount in USD' column had non-numeric characters, which would prevent accurate calculations and analysis. Cleaning the column allowed for reliable financial analysis.

80. What challenges did you face in cleaning the 'Date dd/mm/yyyy' column?

A. The challenges included handling inconsistent date formats and ensuring that invalid or missing dates were properly addressed, so the time-based analysis was not impacted.

81. What does the IQR method do for detecting outliers?

A. The IQR method calculates the range between the first and third quartiles, and identifies values that fall outside the typical range, flagging them as potential outliers.

82. How did you deal with startups that received multiple rounds of funding?

A. Multiple funding rounds for the same startup were aggregated by summing the funding amounts, providing a total funding value for each startup.

83. What were the key industries that attracted the most funding?

A. Industries like Fintech, E-commerce, and Technology received the most funding, reflecting strong investor confidence in these rapidly growing sectors.

84. Why is it essential to standardize column values like 'City Location' and 'Industry Vertical'?

A. Standardizing values ensures consistency and accuracy in analysis, especially when there are multiple variations of city names or industry classifications.

85. How did you use group-by operations in your analysis?

A. Group-by operations were used to aggregate data by categories like year, city, and industry, allowing for the identification of trends and patterns in the funding data.

86. What is the importance of visualizing data distribution using histograms?

A. Histograms help in understanding the spread and concentration of funding amounts, providing insights into how funds are distributed across startups.

87. How did you identify the top-performing cities in terms of funding?

A. The cities were grouped by total funding, and the top 10 cities were identified by sorting them in descending order, revealing the cities attracting the most capital.

88. What does the 'Year' vs. 'Funding Amount' line chart show about funding trends?

A. It reveals the growth or decline of funding over time, highlighting periods of high investor interest and capturing the maturation of the startup ecosystem.

89. How did the visualization of the top industries by funding reveal investor preferences?

A. By visualizing the top-funded industries, it became clear which sectors are receiving the most attention from investors, highlighting growing markets.

90. What did the bar chart showing funding by city tell you about regional startup activity?

A. The bar chart highlighted that startup funding is highly concentrated in a few key cities, with Bengaluru leading, indicating regional disparities in funding access.

91. How do co-investor networks affect the startup funding ecosystem?

A. Co-investor networks enhance the credibility of startups by showing that multiple investors believe in the company, often leading to larger funding rounds.

92. Why was mean imputation used for missing values in 'Amount in USD'?

A. Mean imputation was chosen to replace missing values in 'Amount in USD' to avoid dropping rows and losing valuable data while maintaining reasonable estimates.

93. What insights did you gain from the heatmap of city-industry funding?

A. The heatmap revealed that certain cities dominate specific industries, highlighting areas where startups and funding are concentrated in India.

94. What would happen if we didn't handle missing data in 'Industry Vertical'?

A. Failing to handle missing data in 'Industry Vertical' could lead to biased results and make it difficult to identify trends and patterns in sectoral funding.

95. How did the dataset's cleaning process impact the overall analysis?

A. Data cleaning ensured that only relevant, accurate information was used for analysis, leading to more reliable and actionable insights.

96. What do the 'City Location' and 'Investment Type' features tell you about regional investment trends?

A. They provide insights into how certain cities attract specific investment types, with metropolitan areas often receiving more late-stage funding.

97. How did you visualize the top 10 most funded startups?

A. The startups were grouped by name, and their funding amounts were summed up, followed by a bar chart to visualize which startups attracted the most investment.

98. What was the impact of removing the 'Remarks' column on the analysis?

A. Removing the 'Remarks' column streamlined the dataset and focused analysis on the relevant columns, improving the clarity and focus of insights.

99. How did you track the performance of different sectors over time?

A. By grouping the data by 'Year' and 'Industry Vertical', we tracked the total funding received by each sector and observed the funding trends over time.

100. What do you think about the future of startup funding in India based on the trends in the dataset?

A. Based on the trends, startup funding in India is expected to continue growing, with increasing interest in sectors like Fintech, HealthTech, and E-commerce, especially in key cities like Bangalore.