Startup funding analysis in India

Vishal Sharma

FINAL PROJECT

Abstract

->Dataset used -   the startup funding dataset from kaggle was used. The dataset maybe accessed here -
https://www.kaggle.com/sudalairajkumar/indian-startup-funding

->Research 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?**

Abstract

->Method used: Used numpy , pandas and matplotlib for in-depth analysis of the funding ecosystem of startups in India

->Findings: The findings are as follows:

1. Consumer internet, Technology and E-Commerce industries tends to receive higher number of fundings.Among sub-vertical startups online learning platforms, online pharmacy and food delivery platforms tends to get higher number of startups.
2. Startups located in metro and highly developed cities in India receives higher number of fundings.

Abstract

3. Excluding undisclosed investors, Ratan Tata, Indian Angel Networks and Kalaari Capital makes the most number of fundings to startups

4. The startups in India are mostly funded with Private equity and seed funding type of fundings

Motivation

The goal here is to find out the startups in INDIA which readily receives fundings,the apt location for starting a startup , the prominent fund raisers for startups in INDIA. It is highly useful for an entrepreneur looking to start a sta

Dataset(s)

The dataset consists of the startup names, start date, type of industries they belong to, locations, name of the investor who invested in the startup, type of investment made and the amount of fund provided to start the startup.

The dataset maybe accessed here:

https://www.kaggle.com/sudalairajkumar/indian-startup-funding

Data Preparation and Cleaning

I used pandas functions to check for null values in the dataset, remove the null values and sorted the values. Also, columns unnecessary for analysis were removed.

Research Question(s)

- **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?**

Methods

Pandas was used to prepare and clean data for analysis. Matplotlib was used to create bar graphs showing clear and appropriate relationship between various parameters pertaining to the research questions listed above.
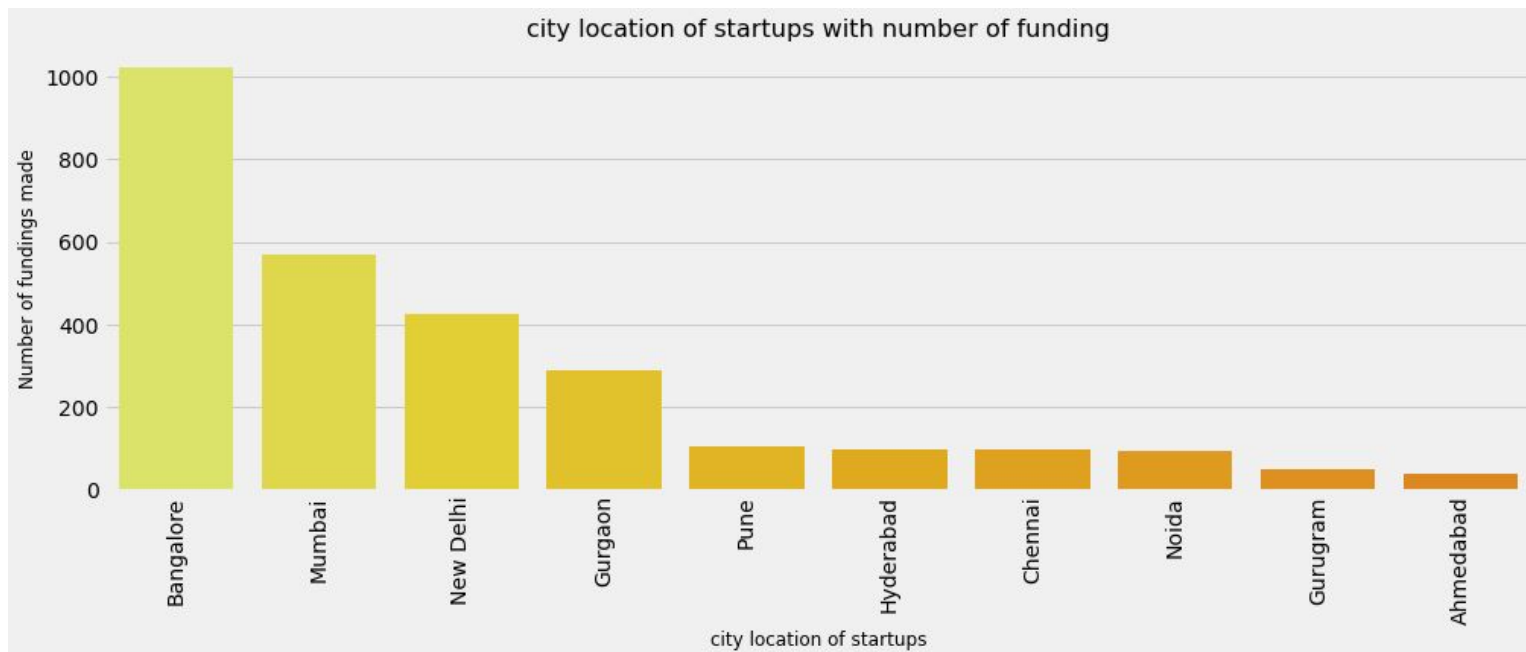
Findings

1. Consumer internet, Technology and E-Commerce industries tends to receive higher number of fundings.Among sub-vertical startups online learning platforms, online pharmacy and food delivery platforms tends to get higher number of startups.

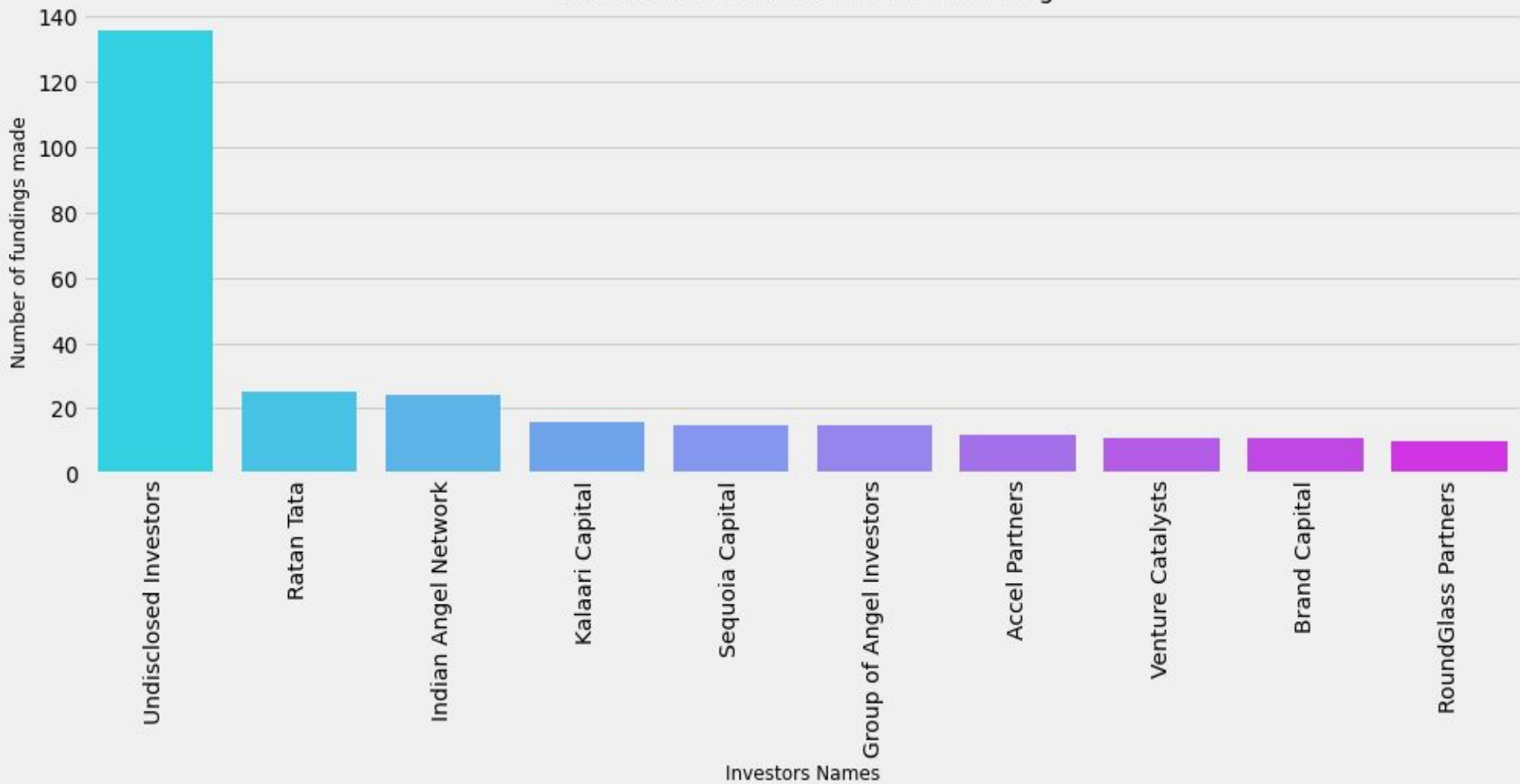Industry vertical of startups with number of funding
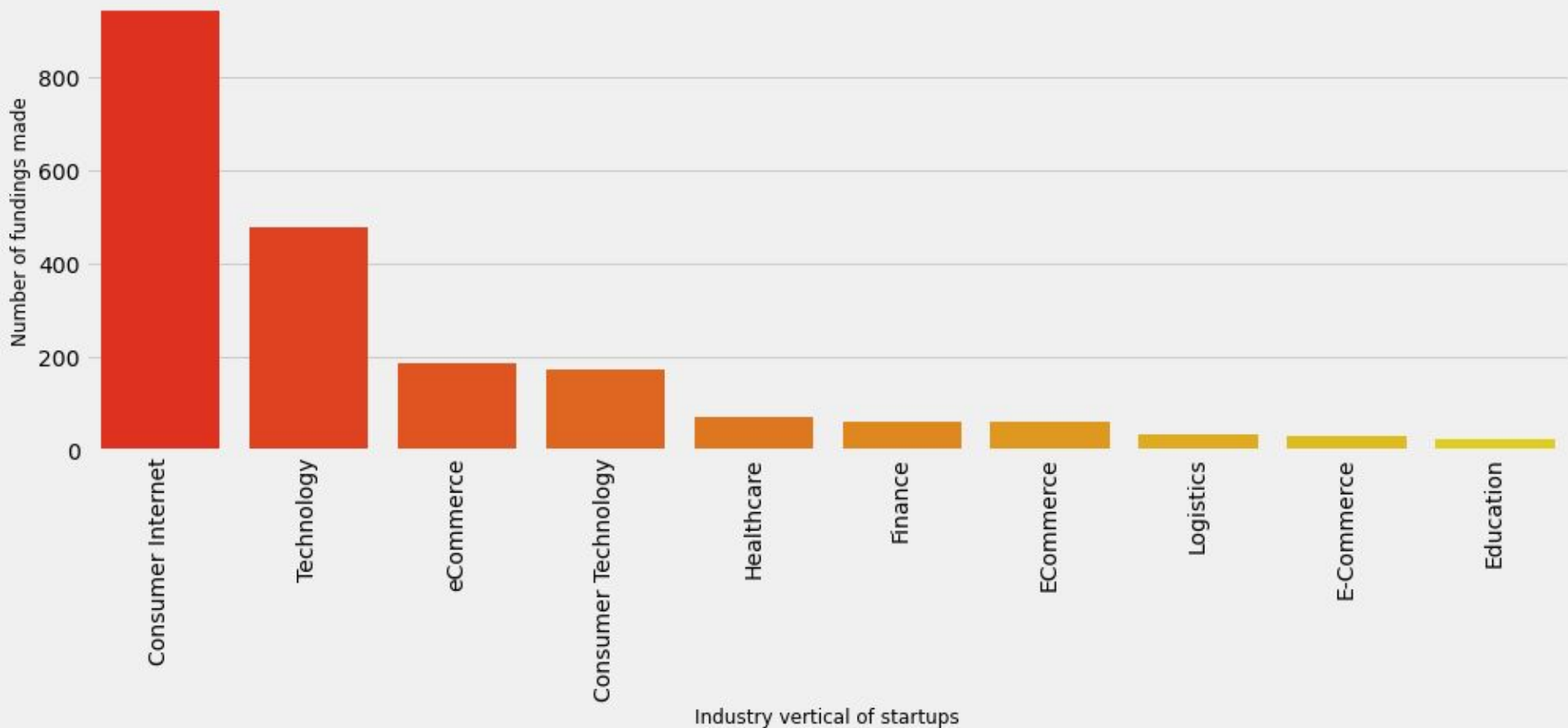
Findings

2. Startups located in metro and highly developed cities in India receives higher number of fundings

Findings

3. Excluding undisclosed investors, Ratan Tata, Indian Angel Networks and Kalaari Capital makes the most number of fundings to startups
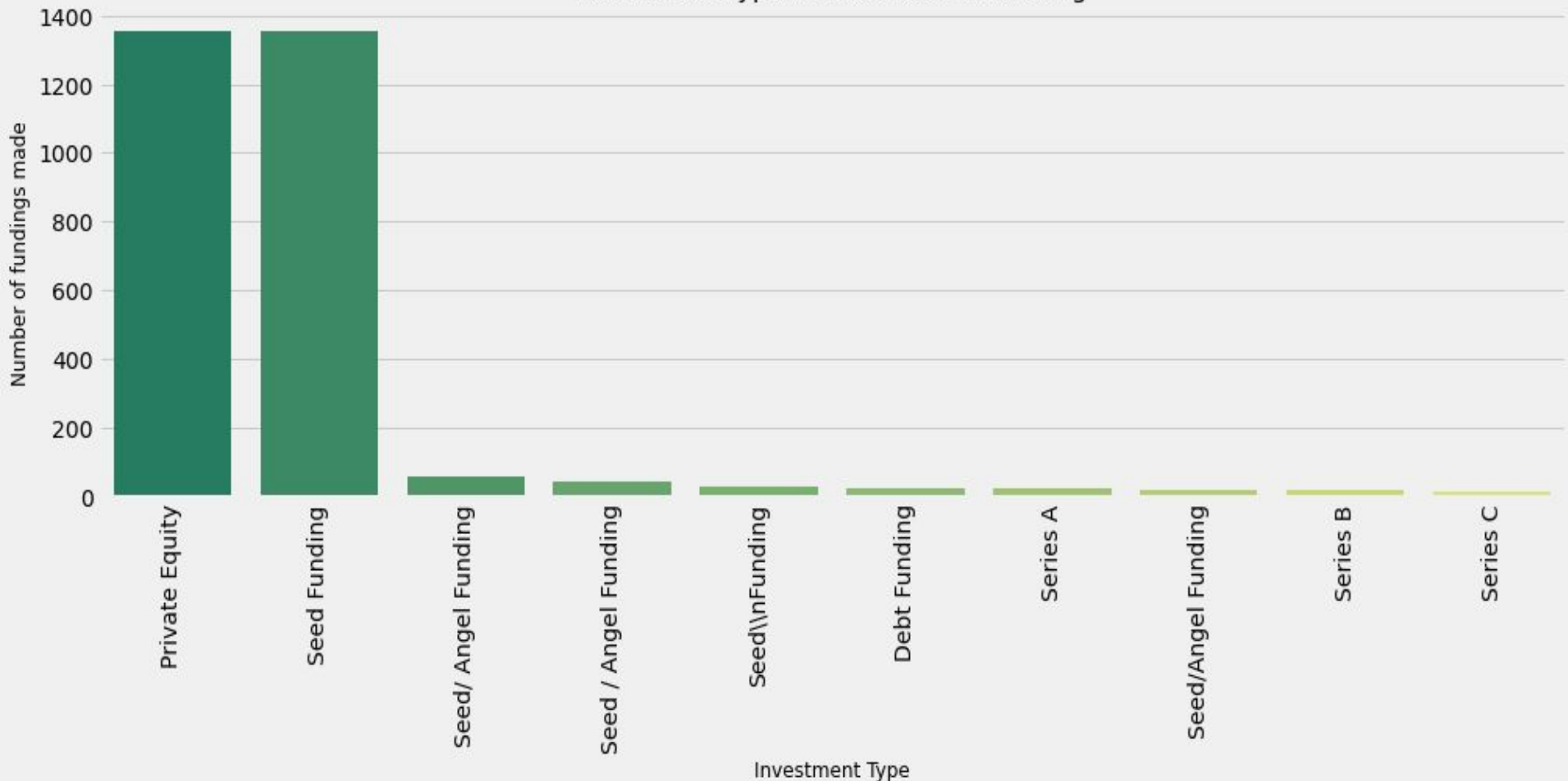
Investors Names with number of funding

Findings

4. The startups in India are mostly funded with Private equity and seed funding type of fundings

Investment Type with number of funding

Limitations

The above analysis is only applicable to start-ups located in India and can't be generalised to the entire globe.

Conclusions

->It will be highly beneficial for a entrepreneur or an enterprise aiming to start a startup in India to invest in Consumer internet, Technology and E-Commerce industries such that it should be located in metro cities such as Bangalore, Delhi etc.

-> They can approach Ratan Tata, Indian Angel Networks and Kalaari Capital for raising funds and choose from Private equity and seed funding whichever they may like .

Acknowledgements

->The data was collected from Kaggle. The dataset maybe accessed here :

   https://www.kaggle.com/sudalairajkumar/indian-startup-funding

-> No i had no one to give any sort of feedback on my work.

References

-> The following notebook from Kaggle was referred for help:

https://www.kaggle.com/monikabishtt/startup-funding-eda

Research 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?**

```python
# for mathematical operations
import numpy as np

# for dataframe operations
import pandas as pd

# for data visualizations
import matplotlib.pyplot as plt
import seaborn as sns

# lets select  a background style
plt.style.use('fivethirtyeight')
plt.rcParams['figure.figsize'] = (16, 7)

# for interactivity
import ipywidgets as widgets
from ipywidgets import interact
from ipywidgets import interact_manual
```

Reading the data

```python
# importing the data
data = pd.read_csv("./startup_funding.csv")
data.head()
```

```python
# changing the names of the columns inside the data
data.columns = ["SNo", "Date", "StartupName", "IndustryVertical", "SubVertical",
                "City", "InvestorsName", "InvestmentType", "AmountInUSD", "Remarks"]

# lets clean the strings
def clean_string(x):
    return str(x).replace("\\xc2\\xa0","").replace("\\\\xc2\\\\xa0", "")

# lets apply the function to clean the data
for col in ["StartupName", "IndustryVertical", "SubVertical", "City",
            "InvestorsName", "InvestmentType", "AmountInUSD", "Remarks"]:
    data[col] = data[col].apply(lambda x: clean_string(x))

# lets check the head of the data
data.head()
```

```python
# checking the column names of the data
data.columns
```

```python
#  checking the shape of the data
print("Size of data", data.shape)
```

Data Cleaning

```python
# missing data

# lets import warnings module
import warnings
warnings.filterwarnings('ignore')

# lets calculate the total missing values in the data
total = data.isnull().sum().sort_values(ascending = False)

# lets calculate the percentage of missing values in the data
percent = ((data.isnull().sum()/data.isnull().count())*100).sort_values(ascending = False)

# lets store the above two values in a dataset called missing data
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent %'])
```

```
# lets check the head of the data
missing_data
```

```
# lets check the values in the Remarks column
data['Remarks'].value_counts()
```

```
# lets remove Remarks column, as it contains a lot of nans, and high cardinal column

data = data.drop(['Remarks'], axis = 1)

# lets check the column names after removing the Remarks Column, as it having
data.columns
```

```
# lets convert the amount column into numerical, so that we can analyze the values inside it

# function to clean the AmounInUsd Column
def clean_amount(x):
    x = ''.join([c for c in str(x) if c in ['0', '1', '2', '3', '4', '5', '6', '7', '8', '9']])
    x = str(x).replace(",","").replace("+","")
    x = str(x).lower().replace("undisclosed","")
    x = str(x).lower().replace("n/a","")
    if x == '':
        x = '-999'
    return x

# lets apply the function on the column
data["AmountInUSD"] = data["AmountInUSD"].apply(lambda x: float(clean_amount(x)))

# lets check the head of the column after cleaning it
plt.rcParams['figure.figsize'] = (15, 3)
data['AmountInUSD'].plot(kind = 'line', color = 'black')
plt.title('Distribution of Amount', fontsize = 15)
plt.show()
```

```
## Cleaning the dates

data['Date'][data['Date']=='12/05.2015'] = '12/05/2015'
data['Date'][data['Date']=='13/04.2015'] = '13/04/2015'
data['Date'][data['Date']=='15/01.2015'] = '15/01/2015'
data['Date'][data['Date']=='22/01//2015'] = '22/01/2015'
data['Date'][data['Date']=='05/072018'] = '05/07/2018'
data['Date'][data['Date']=='01/07/015'] = '01/07/2015'
data['Date'][data['Date']=='\\\\xc2\\\\xa010/7/2015'] = '10/07/2015'
```

How Does the Funding Ecosystem changes with respect to Time?

```
# some of the date formats are written wrongly so we have to clean them thoroughly so that we can analyze these v

# converting them into a Datetime object
data["yearmonth"] = (pd.to_datetime(data['Date'],
        format='%d/%m/%Y').dt.year*100)+(pd.to_datetime(data['Date'],format='%d/%m/%Y').dt.month)

temp = data['yearmonth'].value_counts().sort_values(ascending = False).head(10)
print("Number of funding per month in decreasing order(Top 10)\n",temp)
year_month = data['yearmonth'].value_counts()

# lets plot the data
plt.rcParams['figure.figsize'] = (15, 7)
sns.barplot(year_month.index, year_month.values, palette = 'copper')
plt.xticks(rotation = 90)
plt.xlabel('Year-Month of transaction', fontsize=12)
plt.ylabel('Number of fundings made', fontsize=12)
plt.title("Year-Month Distribution", fontsize=16)
plt.show()
```

What is the General Amount that Startups get in India?

```
# lets check the Maximum funding of a startup
print("Maximum funding to a Startups is : ", data["AmountInUSD"].dropna().sort_values().max())
```

```
# lets check the startups with more than 50crore+ funding

@interact
```

```python
    def check(column = 'AmountInUSD', x = 500000000): # 50 crore funding of startups
        return data[data[column] > x].sort_values(by = 'AmountInUSD', ascending = False)
```

In [ ]:
```python
# lets check out different ventures of Paytm
data[data.StartupName == 'Paytm']
```

In [ ]:
```python
# lets check out different ventures of Flipkart
data[data.StartupName == 'Flipkart']
```

In [ ]:
```python
# lets check the minimum funding in a startup
print("Minimum funding to a Startups is : ", data["AmountInUSD"].dropna().sort_values().min())
```

In [ ]:
```python
# lets check the startups with least funding
data[['AmountInUSD','StartupName']].sort_values(by = 'AmountInUSD', ascending = True).head(5)
```

In [ ]:
```python
data[data['AmountInUSD'] == -999.0].shape
```

In [ ]:
```python
# Average Funding
print("On Average indian startups got funding of : ", data["AmountInUSD"].dropna().sort_values().mean())
```

In [ ]:
```python
# lets check the no. of funding each startsup got

print("Total startups funded : ", len(data["StartupName"].unique()))
print(data["StartupName"].value_counts().head(10))
startupname = data['StartupName'].value_counts().head(20)

# lets plot the data
plt.rcParams['figure.figsize'] = (12, 15)
sns.barplot(y = startupname.index, x = startupname.values, alpha=0.9, palette = 'Dark2')
plt.xticks(rotation='vertical')
plt.xlabel('Startup Name', fontsize=12)
plt.ylabel('Number of fundings made', fontsize=12)
plt.title("Number of fundings Startups Have", fontsize=16)
plt.show()
```

Which Kind of Industries are more preferred for Startups?

In [ ]:
```python
# lets analyze the Industry of the Startups

# data cleaning
data['IndustryVertical'] = data['IndustryVertical'].replace('nan', 'Consumer Technology')

industry = data['IndustryVertical'].value_counts().head(10)
print(industry)

# lets plot the data
plt.rcParams['figure.figsize'] = (15, 5)
sns.barplot(industry.index, industry.values, palette = 'autumn')
plt.xticks(rotation='vertical')
plt.xlabel('Industry vertical of startups', fontsize=12)
plt.ylabel('Number of fundings made', fontsize=12)
plt.title("Industry vertical of startups with number of funding", fontsize=16)
plt.show()
```

In [ ]:
```python
# lets analyze the sub vertical industries

industry = data['SubVertical'].value_counts()[1:].head(20)
print(industry)

# lets plot the data
sns.lineplot(industry.index, industry.values, palette = 'winter')
plt.xticks(rotation='vertical')
plt.xlabel('Subvertical of startups', fontsize=12)
plt.ylabel('Number of fundings made', fontsize=12)
plt.title("Subvertical of startups with number of funding", fontsize=16)
plt.show()
```

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

In [ ]:
```python
# analyzing the effect of Cities on a Startup
```

```
# lets clean the data for better analysis
data['City'] = data['City'].replace(('Bengaluru', 'nan'),('Bangalore', 'Bangalore'))

city = data['City'].value_counts().head(10)
print(city)

# lets plot the data
sns.barplot(city.index, city.values, palette = 'Wistia')
plt.xticks(rotation='vertical')
plt.xlabel('city location of startups', fontsize=12)
plt.ylabel('Number of fundings made', fontsize=12)
plt.title("city location of startups with number of funding", fontsize=16)
plt.show()
```

Who plays the main role in Indian Startups Ecosystem?

In [ ]:
```
from wordcloud import WordCloud

names = data["InvestorsName"][~pd.isnull(data["InvestorsName"])]
wordcloud = WordCloud(max_font_size=50, width=600, height=300, background_color = 'cyan').generate(' '.join(names
plt.figure(figsize=(15,8))
plt.imshow(wordcloud)
plt.title("Wordcloud for Investor Names", fontsize=35)
plt.axis("off")
plt.show()
```

In [25]:
```
# lets analyze the investors on startups

# lets clean the dataset
data['InvestorsName'][data['InvestorsName'] == 'Undisclosed investors'] = 'Undisclosed Investors'
data['InvestorsName'][data['InvestorsName'] == 'undisclosed Investors'] = 'Undisclosed Investors'
data['InvestorsName'][data['InvestorsName'] == 'undisclosed investors'] = 'Undisclosed Investors'
data['InvestorsName'][data['InvestorsName'] == 'Undisclosed investor'] = 'Undisclosed Investors'
data['InvestorsName'][data['InvestorsName'] == 'Undisclosed Investor'] = 'Undisclosed Investors'
data['InvestorsName'][data['InvestorsName'] == 'Undisclosed'] = 'Undisclosed Investors'
data['InvestorsName'][data['InvestorsName'] == 'nan'] = 'Undisclosed Investors'

# lets check the value counts
investors = data['InvestorsName'].value_counts().head(10)
print(investors)

# lets plot the data
sns.barplot(investors.index, investors.values, palette = 'cool')
plt.xticks(rotation='vertical')
plt.xlabel('Investors Names', fontsize=12)
plt.ylabel('Number of fundings made', fontsize=12)
plt.title("Investors Names with number of funding", fontsize=16)
plt.show()
```

```
Undisclosed Investors       136
Ratan Tata                   25
Indian Angel Network         24
Kalaari Capital              16
Sequoia Capital              15
Group of Angel Investors     15
Accel Partners               12
Venture Catalysts            11
Brand Capital                11
RoundGlass Partners          10
Name: InvestorsName, dtype: int64
```
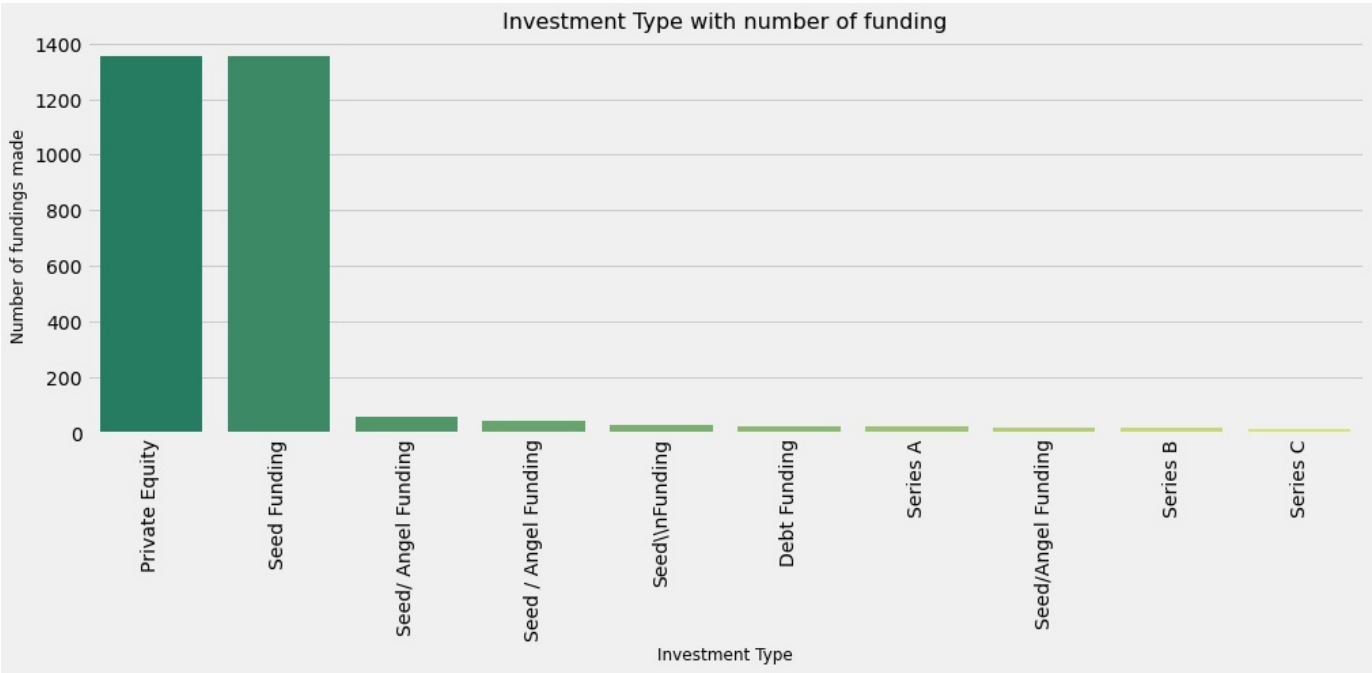
Undisclosed

Indian Ange

Kala:

Sequc

Group of Angel

Acce

Venture

Bra

RoundGlas

Investors Names
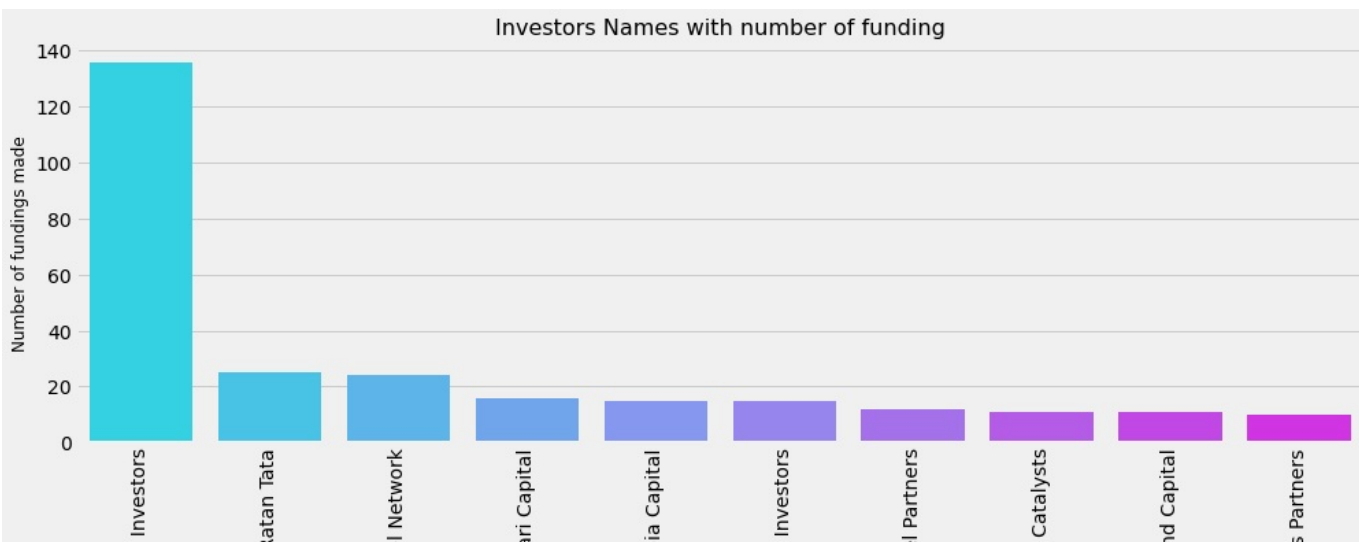
What are the different Types of Funding for Startups?

In [26]:
```python
# lets analyze the investment
investment = data['InvestmentType'].value_counts().head(10)
print(investment)

# lets clean the dataset
data['InvestmentType'][data['InvestmentType'] == 'SeedFunding'] = 'Seed Funding'
data['InvestmentType'][data['InvestmentType'] == 'Crowd funding'] = 'Crowd Funding'
data['InvestmentType'][data['InvestmentType'] == 'PrivateEquity'] = 'Private Equity'

# lets plot the data
sns.barplot(investment.index, investment.values, palette = 'summer')
plt.xticks(rotation='vertical')
plt.xlabel('Investment Type', fontsize=12)
plt.ylabel('Number of fundings made', fontsize=12)
plt.title("Investment Type with number of funding", fontsize=16)
plt.show()
```

```
Private Equity          1356
Seed Funding            1355
Seed/ Angel Funding       60
Seed / Angel Funding      47
Seed\nFunding             30
Debt Funding              25
Series A                  24
Seed/Angel Funding        23
Series B                  20
Series C                  14
Name: InvestmentType, dtype: int64
```


In [ ]:

In [ ]:

In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js