 Mahatma Education Society’s



**Pillai College of Arts, Commerce & Science**

**(Autonomous)**

**Affiliated to University of Mumbai**

NAAC Accredited 'A' grade (3 cycles)

Best College Award by University of Mumbai

ISO 9001:2015 Certified

**CERTIFICATE**

*This is to certify that Mr. /Miss.****\_\_ANCHAL SINGH\_\_\_\_*** *of* ***S.Y B.Sc. I.T. Semester lV*** *has completed the project work in the Subject of****Data Science*** *during the academic year 2021-22 under the guidance of Prof****. \_\_Sanjana Bhangale\_\_\_****being the partial requirement for the fulfillment of the curriculum of* ***Degree of  Bachelor of Science in Information Technology****,* ***University of Mumbai****.*

*Place: Panvel*

*Date:*

*Name & Signature of faculty*

DATA SCIENCE PROJECT

Name: Anchal Singh

Roll no:7113

Class: SYBSCIT

Div: B

**DATASET :STARTUP\_FUNDING**

**Context**

Interested in the Indian startup ecosystem just like me? Wanted to know what type of startups are getting funded in the last few years? Wanted to know who are the important investors? Wanted to know the hot fields that get a lot of funding these days?  
This dataset is a chance to explore the Indian start up scene. Deep dive into funding data and derive insights into the future!

**Content**

This dataset has funding information of the Indian startups from January 2015 to August 2017. It includes columns with the date funded, the city the startup is based out of, the names of the funders, and the amount invested (in USD).

**Inspiration**

Possible questions which could be answered are:

* How does the funding ecosystem change with time?
* Do cities play a major role in funding?
* Which industries are favored by investors for funding?
* Who are the important investors in the Indian Ecosystem?
* How much funds does startups generally get in India?

**Columns:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |

1. Sr No
2. Date dd/mm/yyyy
3. Startup Name
4. Industry Vertical
5. SubVertical
6. City Location
7. Investors Name
8. InvestmentnType
9. Amount in USD
10. Remarks

**IMPORTING THE LIBRARIES**

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import pandas as pd

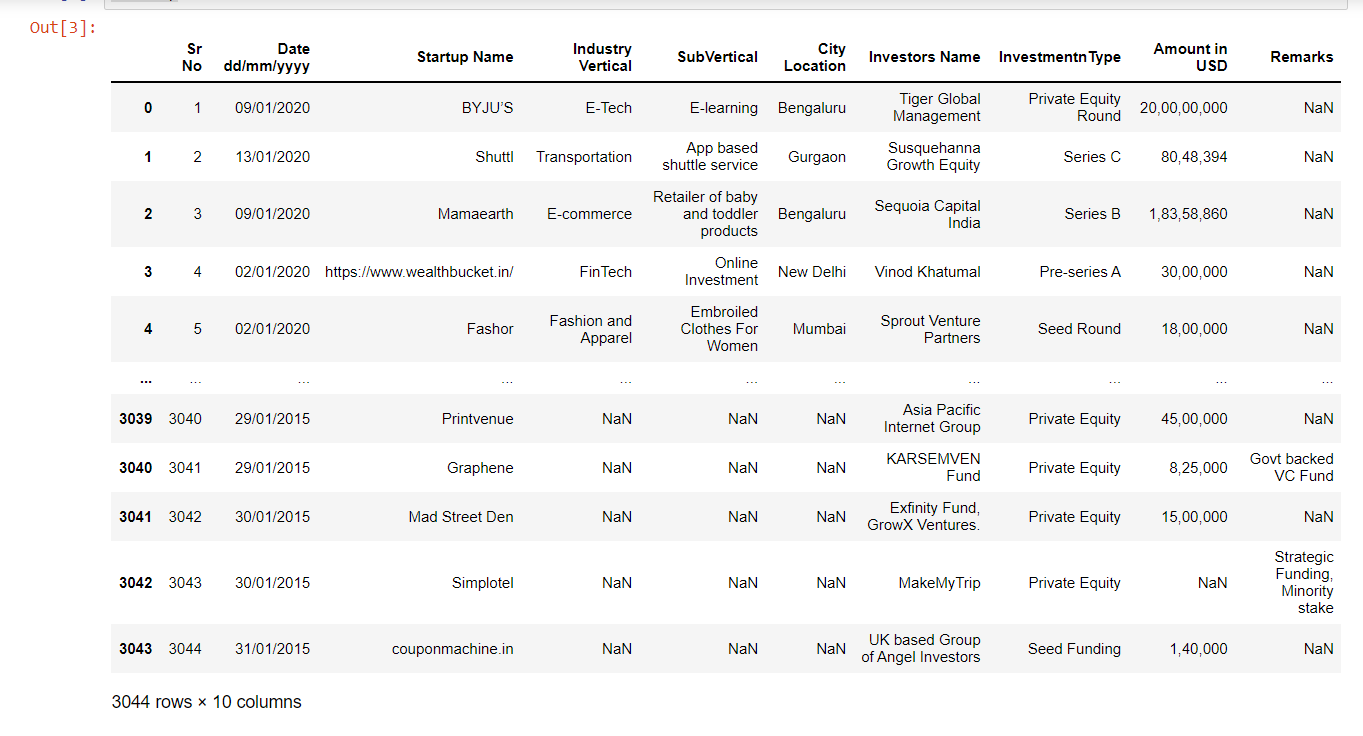
import plotly

import plotly.express as px

**READING THE DATA**

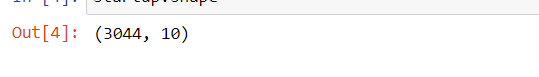
startup=pd.read\_csv('file:///E:\PILLAI%20COLLEGE\SECOND%20YEAR%20SEM%204\DATA%20SCIENCE\startup\_funding.csv')

startup



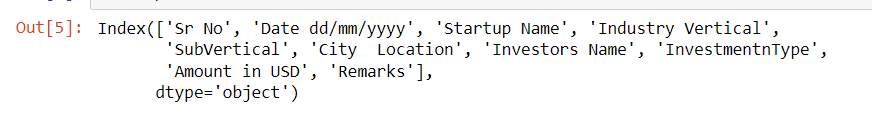
**NUMBER OF ROWS AND COLUMNS**

startup.shape



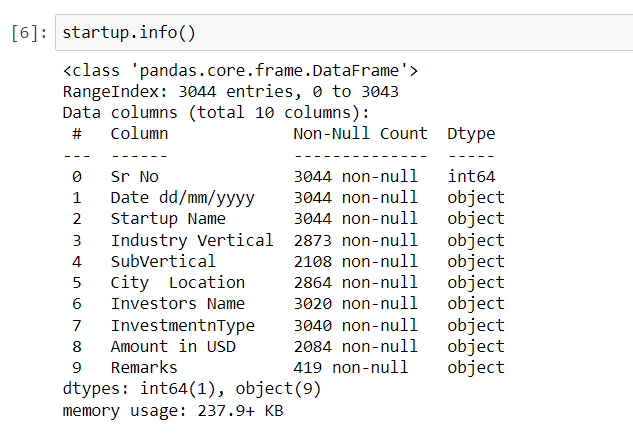
**COLUMNS**

startup.columns



**CONCISE SUMMARY OF A DATASET**

startup.info()



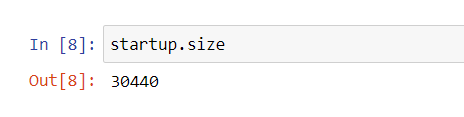
**DIMENSION OF THE DATA**

startup.ndim



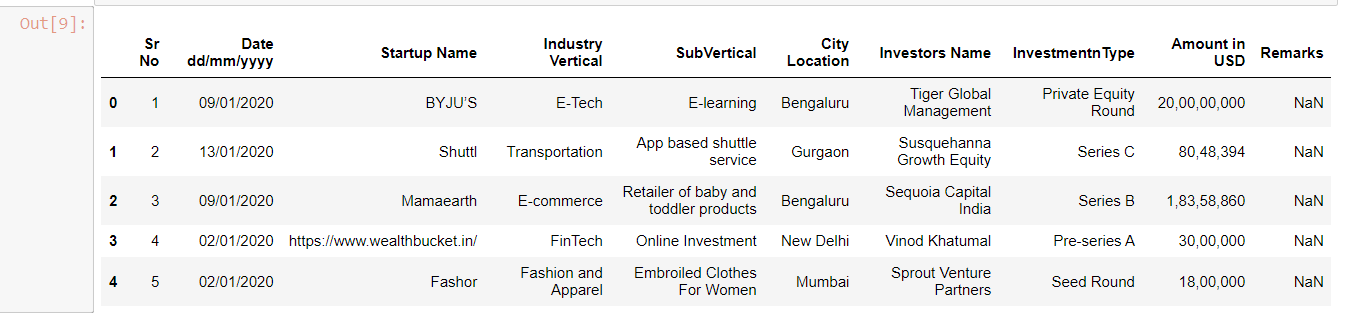
**SIZE**

startup.size



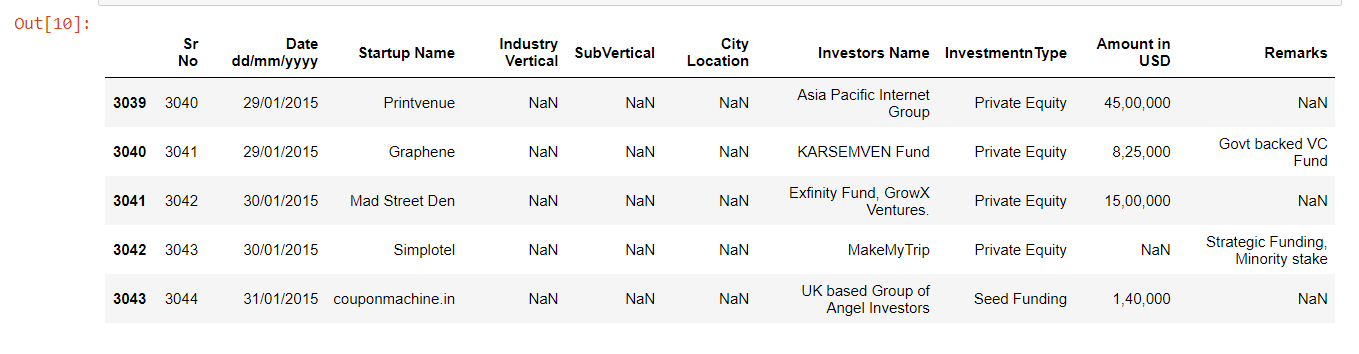
**DISPLAYING FIRST 5 ROWS**

startup.head()



**DISPLAYING LAST 5 ROWS**

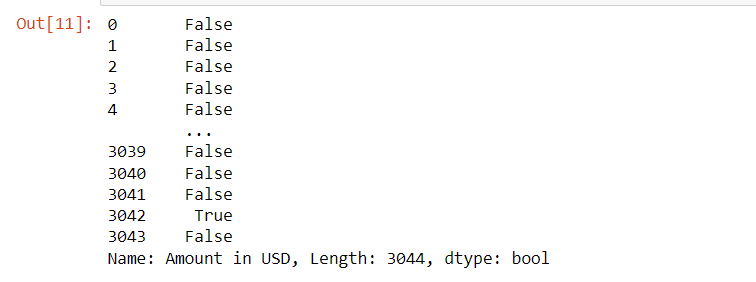
startup.tail()



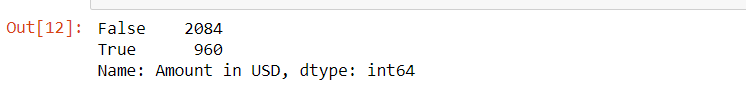
**DATA CLEANING**

**CHECKING NULL VALUES IN AMOUNT COLUMN**

startup['Amount in USD'].isnull()

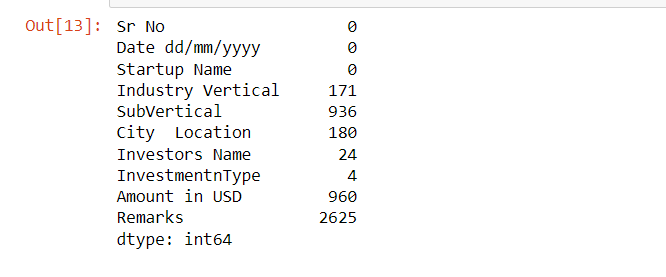


startup['Amount in USD'].isnull().value\_counts()



**CHECKING TOTAL NULL VALUES IN EVERY COLUMN**

startup.isnull().sum()

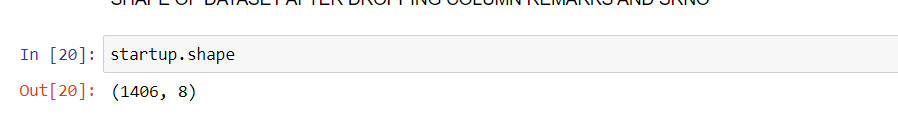


**DROPPING COLUMN REMARKS AND SRNO**

startup=startup.drop(["Remarks","Sr No"],axis=1)

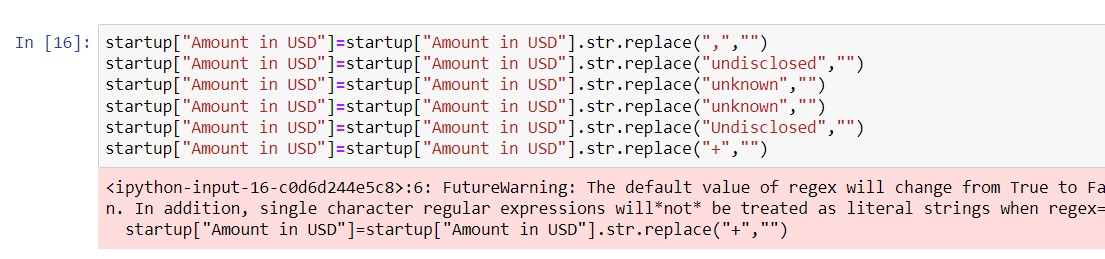
**SHAPE OF DATASET AFTER DROPPING COLUMN REMARKS AND SRNO**

startup.shape



startup=startup.dropna()

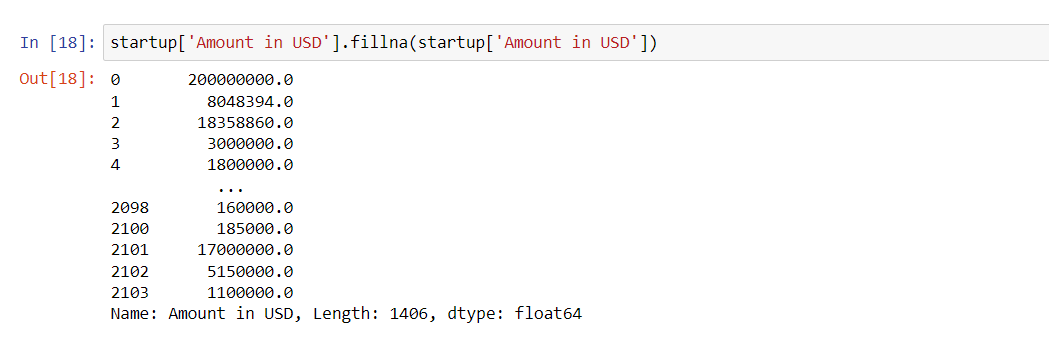
**CONVERTING AMOUNT IN USD FROM STRING TO INTEGER**



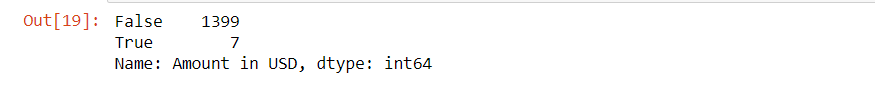
startup["Amount in USD"]=pd.to\_numeric(startup["Amount in USD"])

**FILLING NULL VALUE IN COLUMN Amount in USD WITH AVERAGE VALUE OF COLUMN**

startup['Amount in USD'].fillna(startup['Amount in USD'])

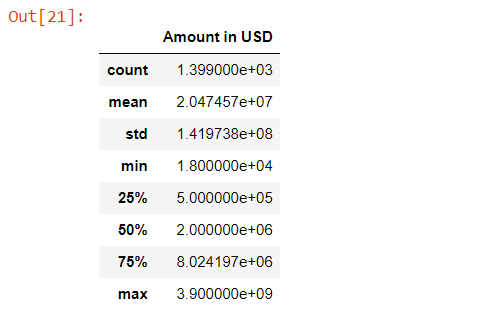


startup['Amount in USD'].isnull().value\_counts()



**VIEWING SOME BASIC STATISTICAL DETAILS LIKE PERCENTILE, MEAN, STD ETC**

startup.describe()

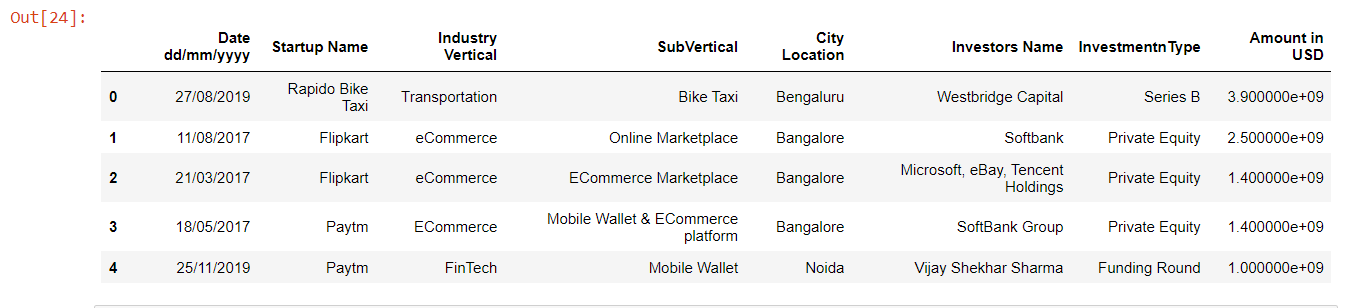


**VIEWING AMOUNT IN USD COLUMN IN DECREASING AMOUNT**

startup=startup.sort\_values(by=["Amount in USD"],ascending=False).reset\_index()

startup=startup.drop(["index"],axis=1)

startup.head()

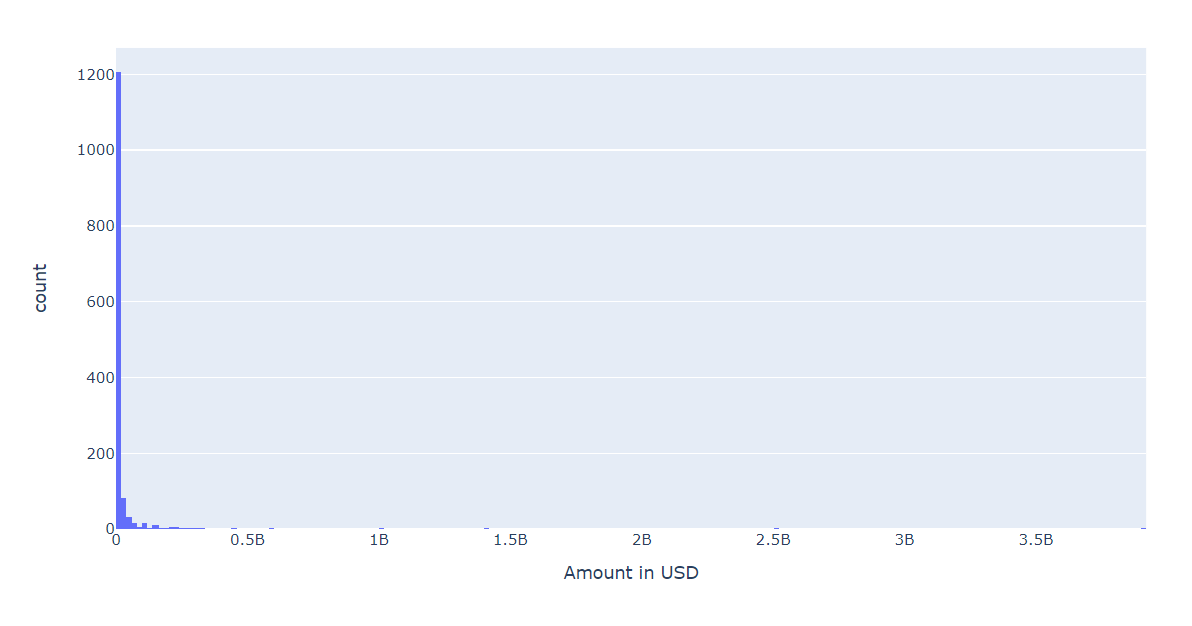


**LETS CHECK THE FUNDING RANGE PROVIDED BY INVESTORS**

plt.figure(figsize=(15,8))

sns.set(style="darkgrid")

px.histogram(startup, x="Amount in USD")



The maximum density of fundings provided by invester lies in 0-0.5B and ``Rapido Bike Tax'' is the only startup who has recieved funding of 3.9B

**LETS CHECK THE STARTUPS WHO HAS RECIEVED BOTTOM 20 MINOR FUNDINGS**

startup=startup.sort\_values(by=["Amount in USD"],ascending=True).reset\_index()

start\_botfun=startup[0:20]

start\_botfun[["Startup Name","Amount in USD"]]

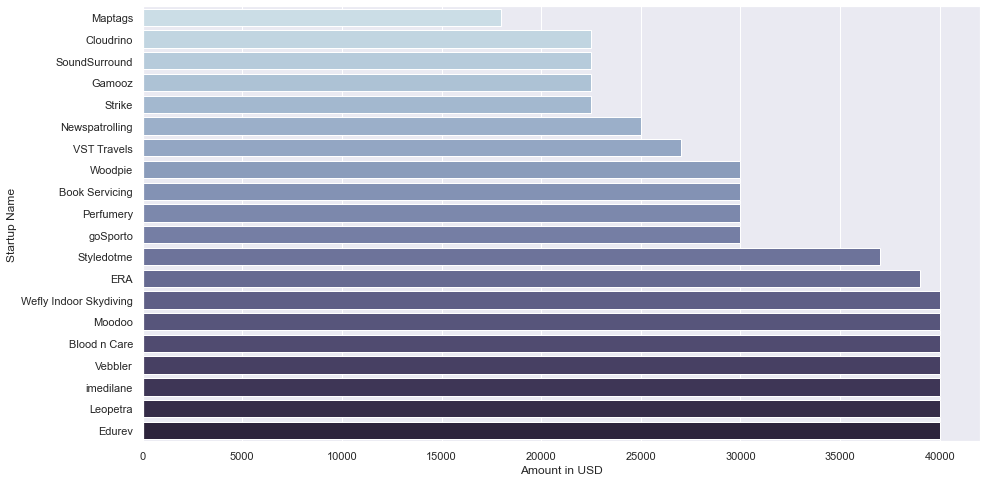


plt.figure(figsize=(15,8))

sns.set(style="darkgrid")

sns.barplot(data=start\_botfun,y="Startup Name",x="Amount in USD",palette="ch:s=.25,rot=-.25")

----



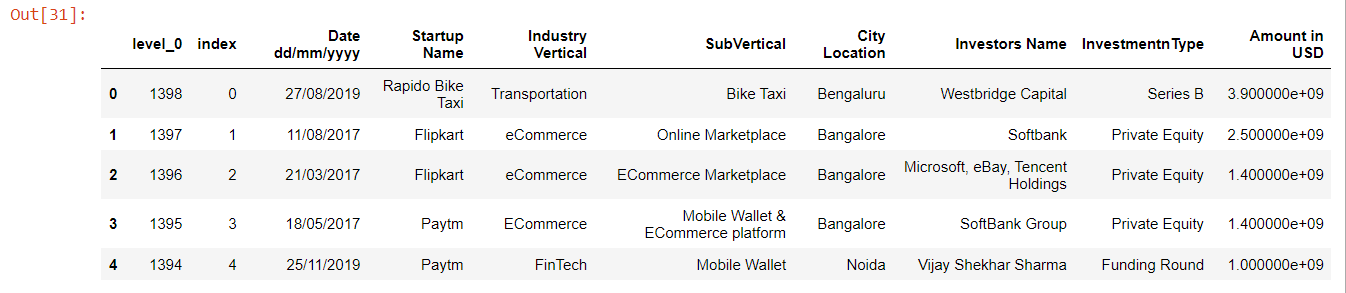
Startups like Maptags, Cloudrino, SoundSurround have received Lowest

fundings

**LETS CHECK THE STARTUPS WHO HAS RECIEVED TOP 20 MAJOR** **FUNDINGS**

startup=startup.sort\_values(by=["Amount in USD"],ascending=False).reset\_index()

startup.head()

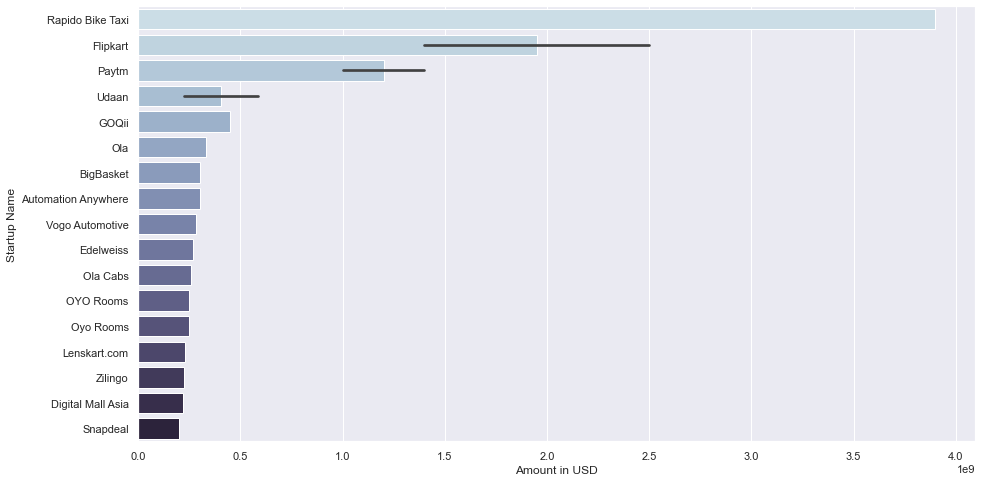
 start\_topfun=startup[0:20]

plt.figure(figsize=(15,8))

sns.set(style="darkgrid")

sns.barplot(data=start\_topfun,y="Startup Name",x="Amount in USD",palette="ch:s=.25,rot=-.25")

----



Startups like Rapido bike taxi, Flipkart, Udaan and Paytm have received

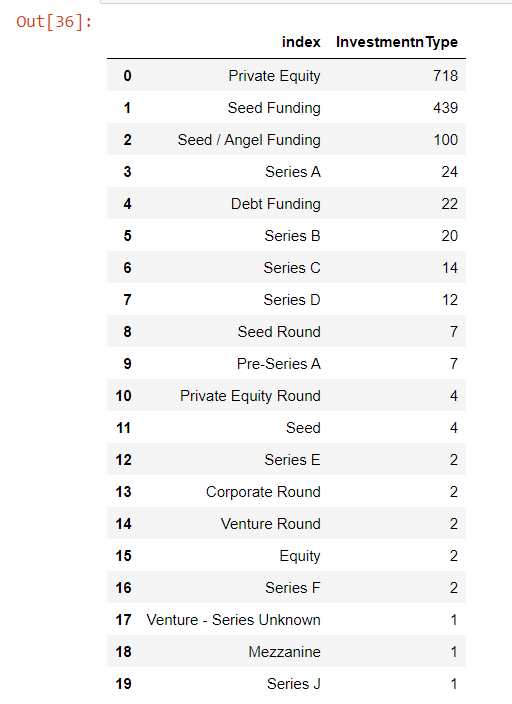
major fundings

**LETS CHECK WHAT TYPE OF INVESTMENT ARE THE INVESTOR REALLY INTERESTED IN**



start\_INtype=startup["InvestmentnType"].value\_counts()[0:20].reset\_index()

start\_INtype



----

plt.figure(figsize=(15,12))

sns.set(style="darkgrid")

sns.barplot(data=start\_INtype,y="index",x="InvestmentnType",palette="ch:s=.25,rot=-.25")

plt.title("Investment Type in which Investor are really intrested")

plt.xlabel("count")

plt.ylabel("Investment Type")

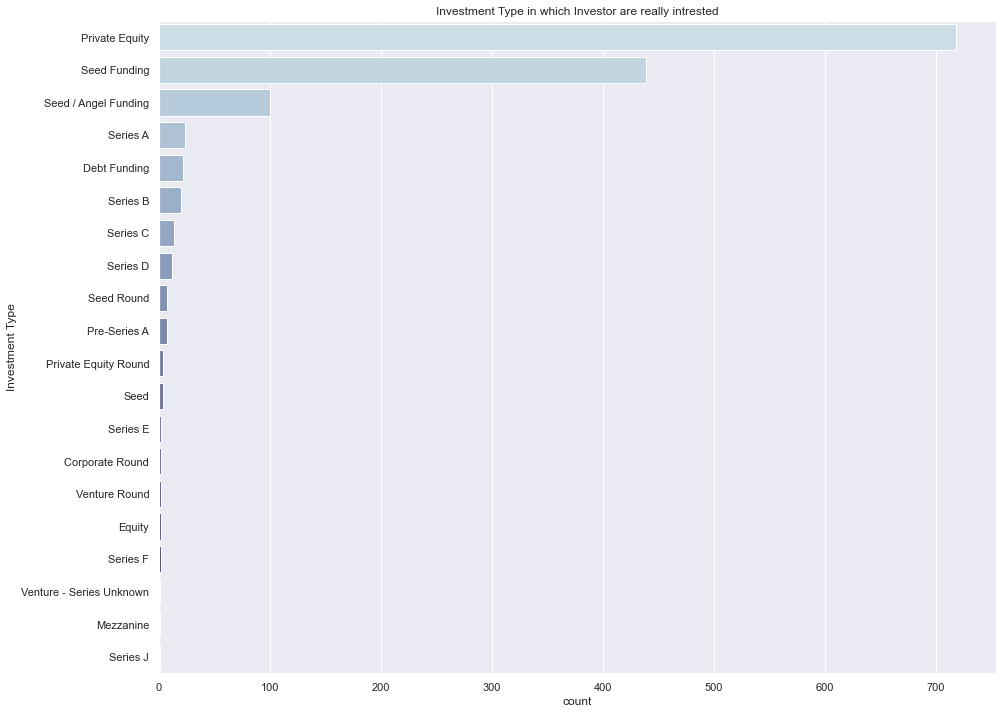
plt.show()

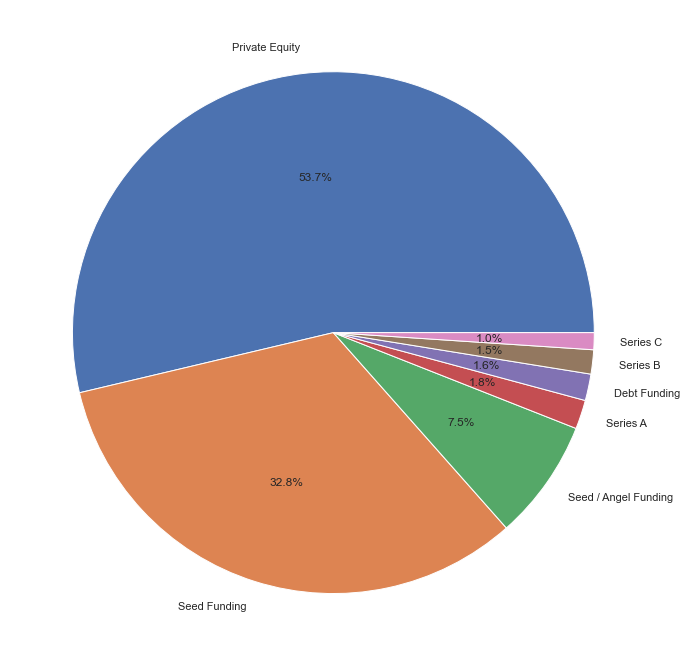
plt.figure(figsize=(15,12))

plt.pie(start\_INtype["InvestmentnType"][0:7],labels=start\_INtype["index"][0:7],autopct="%0.1f%%")

plt.show()

----



----

Maximum investor are intrested in Private equity, Seed funding and Seed / Angel Funding as their Investment Type

**LETS CHECK THE INDUSTRY VERTICAL IN WHICH INVESTOR ARE INTERESTED**

start\_indver=startup["Industry Vertical"].value\_counts().reset\_index()[0:10]

start\_indver



plt.figure(figsize=(15,12))

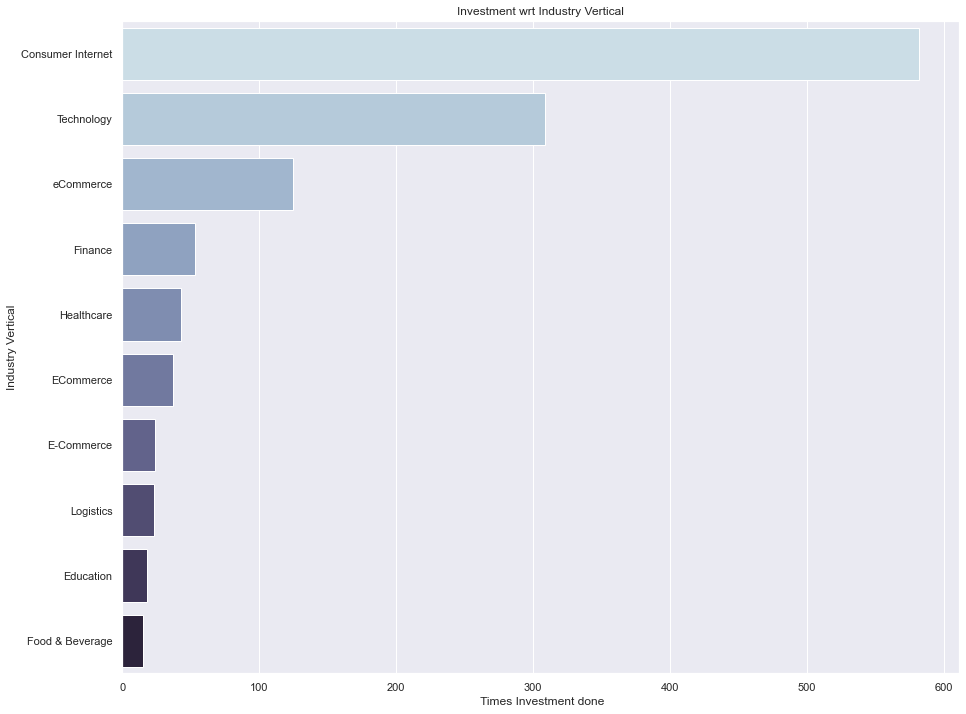
sns.set(style="darkgrid")

sns.barplot(data=start\_indver,y="index",x="Industry Vertical",palette="ch:s=.25,rot=-.25")

plt.title("Investment wrt Industry Vertical")

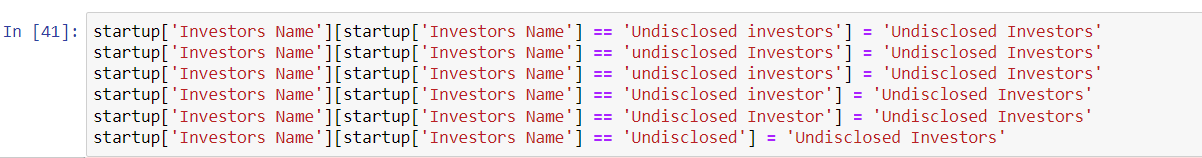
plt.xlabel(" Times Investment done")

plt.ylabel("Industry Vertical")

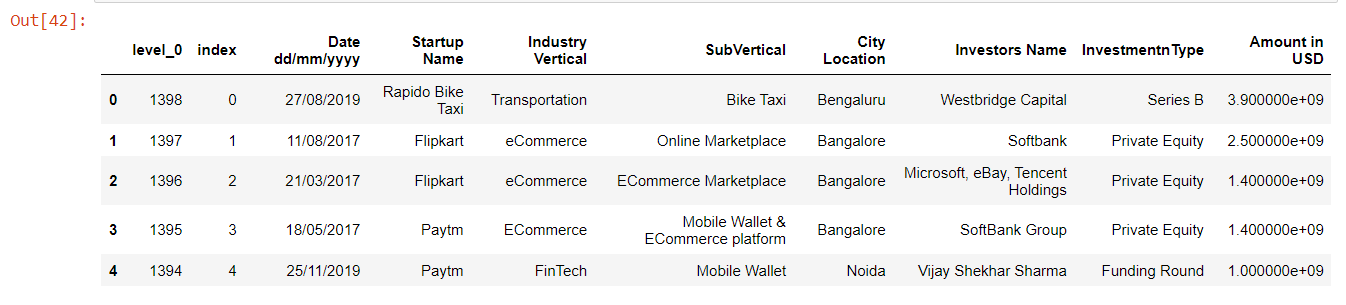


Investor have been funding Consumer Internet and Technology

**NOW LETS CHECK WHO ARE THE TOP 20 INVESTORS EXCLUDING THE UNDISCLOSED INVESTOR**



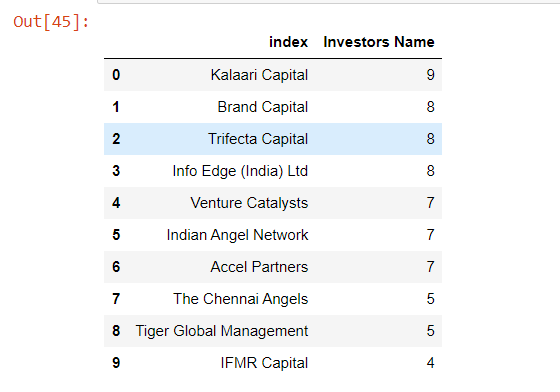
startup.head()



start\_un=startup[startup["Investors Name"]!="Undisclosed Investors"]

start\_inves=start\_un["Investors Name"].value\_counts()[0:10].reset\_index()

start\_inves



plt.figure(figsize=(15,12))

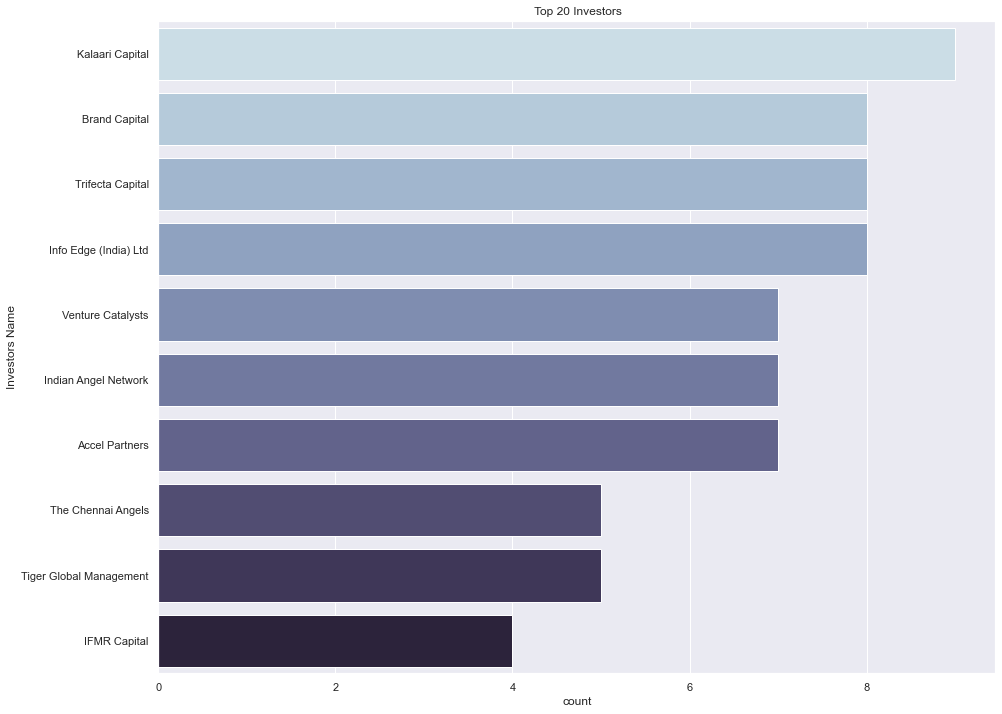
sns.set(style="darkgrid")

sns.barplot(data=start\_inves,y="index",x="Investors Name",palette="ch:s=.25,rot=-.25")

plt.title(" Top 20 Investors")

plt.xlabel("count")

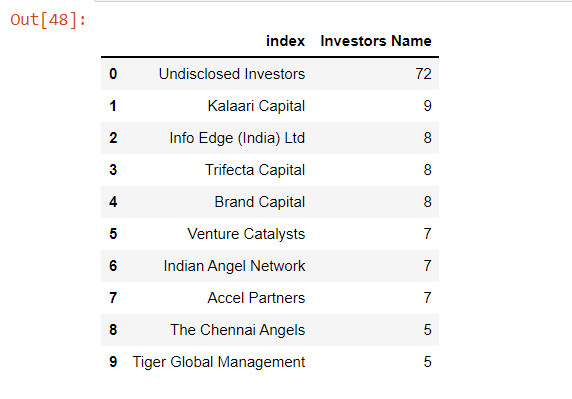
plt.ylabel("Investors Name")



**NOW LETS CHECK WHO ARE THE TOP 20 INVESTORS INCLUDING THE UNDISCLOSED INVESTOR**

start\_invun=startup["Investors Name"].value\_counts()[0:10].reset\_index()

start\_invun



plt.figure(figsize=(15,12))

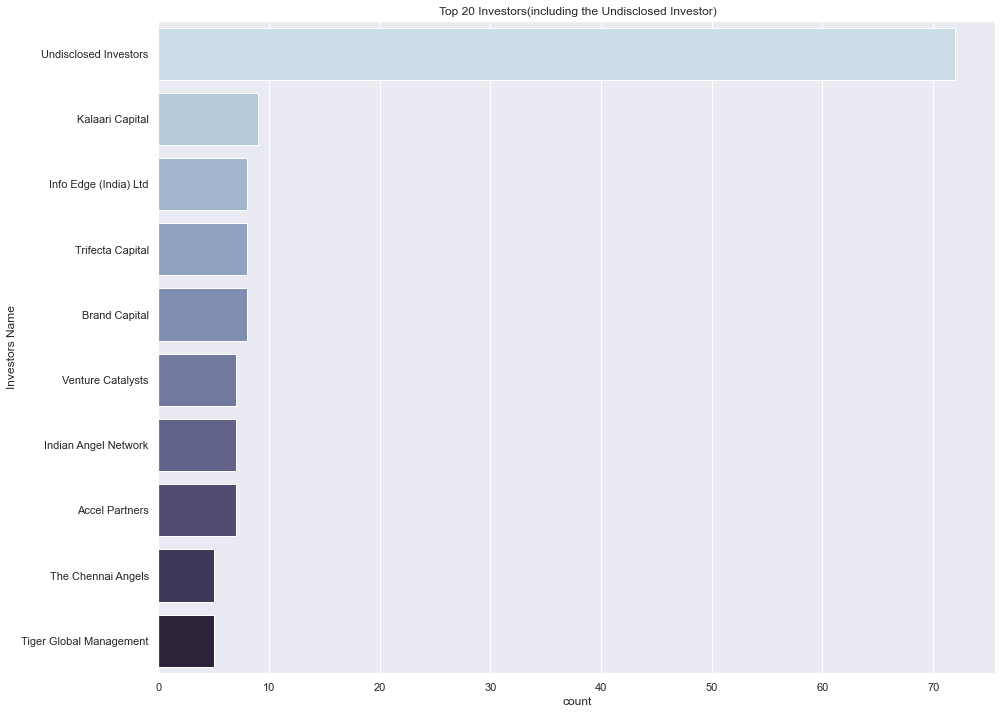
sns.set(style="darkgrid")

sns.barplot(data=start\_invun,y="index",x="Investors Name",palette="ch:s=.25,rot=-.25")

plt.title(" Top 20 Investors(including the Undisclosed Investor)")

plt.xlabel("count")

plt.ylabel("Investors Name")



**IN WHICH CITY ARE THE HIGHEST NUMBER OF STARTUPS LOCATED**

removes null values from entire dataset (column city ,location)

data\_cities=startup.dropna(how="any",axis=0,subset=['City Location'])

data\_cities.shape

data\_cities\_sum=pd.DataFrame(data\_cities.groupby(['City Location']).sum()["Amount in USD"])

data\_cities\_sum['% Investements']=((data\_cities\_sum['Amount in USD']/data\_cities\_sum['Amount in USD'].sum())\*100).round(2)

data\_cities\_sum=data\_cities\_sum.reset\_index().sort\_values(by='% Investements',ascending=False)

Top\_10 = data\_cities\_sum[['City Location','% Investements']][0:10]

Top\_10.loc[len(Top\_10.index)] = ['Others', 100-(data\_cities\_sum['% Investements'][0:10]).sum()]

Top\_10=Top\_10.sort\_values(by='% Investements',ascending=False)

plt.figure(figsize=(6,6))

plt.title("Distribution of Investments by City")

colors1 = ['#191970','#001CF0','#0038E2','#0055D4','#0071C6','#008DB8','#00AAAA','#00C69C','#00E28E','#00FF80','#4EF4D1']

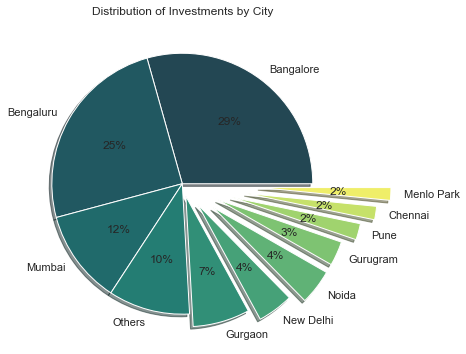
colors=['#234753','#215861','#206A6B','#247D73','#318F77','#46A178','#60B276','#7EC372','#A0D36E','#C6E16A','#EFEE69']

plt.axis("equal")

g3=plt.pie(Top\_10['% Investements'],labels=Top\_10['City Location'],colors=colors,autopct='%0.0f%%',

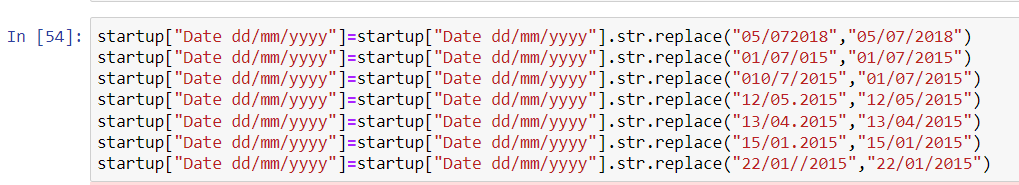
shadow=True,explode=[0, 0, 0, 0, 0.1, 0.2,0.3,0.3,0.4,0.5,0.6])

plt.show()



**HOW IS THE FUNDING ECOSYSTEM EVOLVING WITH TIME ?**

**CONVERTING DATE TO DATETIME AND EXTRACTING MONTH AND YEAR VALUES**



startup['Date dd/mm/yyyy']=pd.to\_datetime(startup['Date dd/mm/yyyy'])

startup['Year']=startup['Date dd/mm/yyyy'].dt.year

startup['Month']=startup['Date dd/mm/yyyy'].dt.month

startup['YearMonth']=startup['Year']\*100+startup['Month']

data\_time=startup[['Date dd/mm/yyyy','YearMonth','City Location','Amount in USD']]

data\_time=data\_time.dropna(how="any",axis=0,subset=['City Location'])

data\_time=pd.DataFrame(data\_time[['YearMonth','City Location']].value\_counts()).reset\_index().sort\_values(by=['City Location','YearMonth'],ascending=True)

data\_time=data\_time[data\_time['City Location'].isin(['Bengaluru','Mumbai','New Delhi','Gurgaon','Noida'])]

data\_time.columns=['YearMonth','City Location','Count']

**catplot: relationship between numerical variable and one or more**

**categorical variables**

color3=['#50F9F1','#6AE5A8','#96C96A','#B7AA47','#C78845']

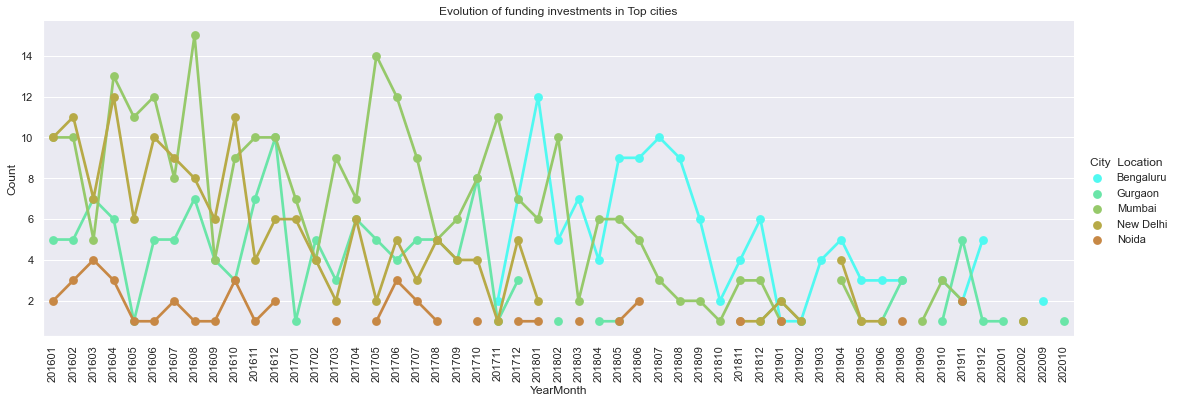
sns.set\_palette(sns.color\_palette(color3))

g5=sns.catplot(data=data\_time,x='YearMonth',y='Count',hue='City Location',kind="point",height=5, aspect=3.0)

g5.set(title='Evolution of funding investments in Top cities')

plt.xticks(rotation=90)

plt.show()



**NUMBER OF FUNDING PER MONTH IN DECREASING ORDER**

startup["yearmonth"] = (pd.to\_datetime(startup['Date dd/mm/yyyy'],

format='%d/%m/%Y').dt.year\*100)+(pd.to\_datetime(startup['Date dd/mm/yyyy'],format='%d/%m/%Y').dt.month)

temp = startup['yearmonth'].value\_counts().sort\_values(ascending = False).head(10)

print("Number of funding per month in decreasing order(Top 10)\n",temp)

year\_month = startup['yearmonth'].value\_counts()

# lets plot the data

plt.rcParams['figure.figsize'] = (15, 7)

sns.barplot(year\_month.index, year\_month.values, palette = 'copper')

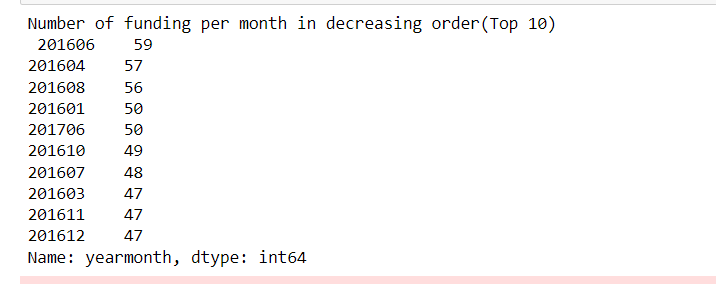
plt.xticks(rotation = 90)

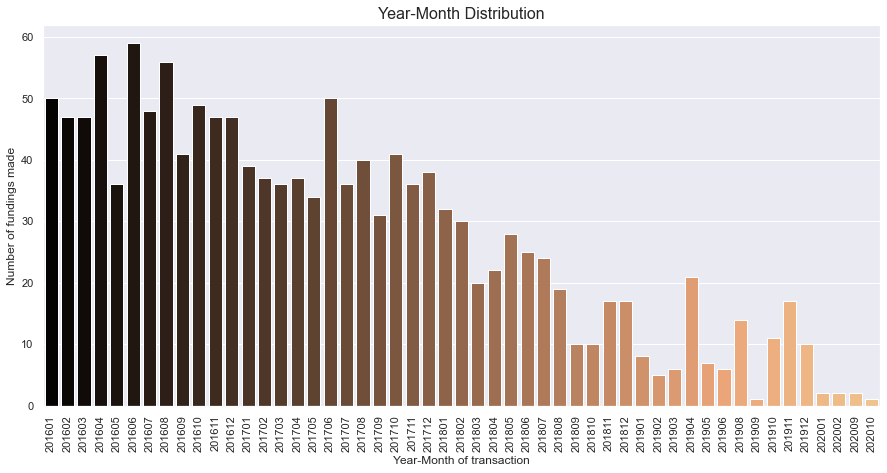
plt.xlabel('Year-Month of transaction', fontsize=12)

plt.ylabel('Number of fundings made', fontsize=12)

plt.title("Year-Month Distribution", fontsize=16)

plt.show()





**AVERAGE FUNDING**

print("On Average indian startups got funding of : ", startup["Amount in USD"].dropna().sort\_values().mean())

On Average indian startups got funding of : 20474571.147248033

**LETS CHECK THE MAXIMUM FUNDING OF A STARTUP**

print("Maximum funding to a Startups is : ", startup["Amount in USD"].dropna().sort\_values().max())

Maximum funding to a Startups is : 3900000000.0

**LETS CHECK THE NO. OF FUNDING EACH STARTSUP GOT**

print("Total startups funded : ", len(startup["Startup Name"].unique()))

print(startup["Startup Name"].value\_counts().head(10))

StartupName = startup['Startup Name'].value\_counts().head(20)

# lets plot the data

plt.rcParams['figure.figsize'] = (8, 6)

sns.barplot(x = StartupName.index, y = StartupName.values, alpha=0.9, palette = 'Dark2')

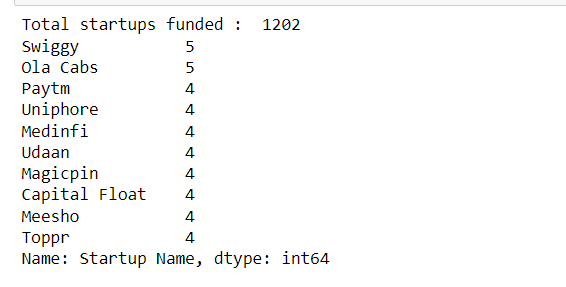
plt.xticks(rotation='vertical')

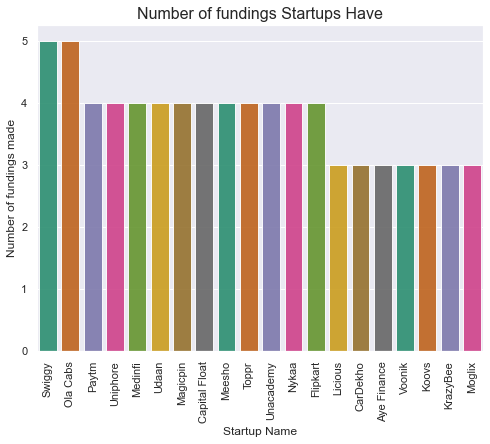
plt.ylabel('Number of fundings made', fontsize=12)

plt.xlabel('Startup Name', fontsize=12)

plt.title("Number of fundings Startups Have", fontsize=16)

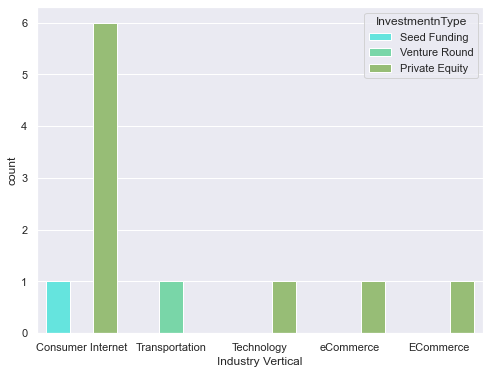
plt.show()





**NUMBER OF OCCURRENCES OF INDUSTRY VERTICAL BASED ON A INVESTMENTNTYPE OF CATEGORY.**

ax = sns.countplot(x="Industry Vertical", hue="InvestmentnType",data=startup.value\_counts().head(11).reset\_index())



**UNDERSTANDING INVESTMENT TYPE PER INDUSTRY VERTICAL ANS AMOUNT IN USD**

pivot = startup.value\_counts().head(11).reset\_index().pivot\_table(

index=["Industry Vertical"],

columns=["InvestmentnType"],

values="Amount in USD",

aggfunc=np.average)

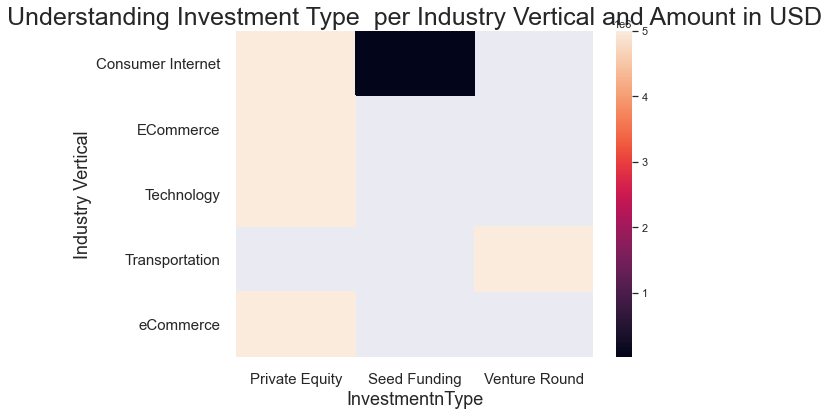
p=sns.heatmap(pivot)

p.set\_title("Understanding Investment Type per Industry Vertical and Amount in USD",fontsize=25)

plt.tick\_params(axis='both',labelsize=15,pad=10)

plt.xlabel('InvestmentnType',fontsize=18)

plt.ylabel('Industry Vertical',fontsize=18)



**CONCLUSION:** Using STARTUP\_FUNDING dataset we were able to plot different plots and compare various data ,display various information.

Succesfully plotted visualised data using matplotlib,seaborn,pandas,numpy libraries.