Code

* [Show All Code](#gjdgxs)
* [Hide All Code](#gjdgxs)
* [Download Rmd](#gjdgxs)

# R Notebook

Introduction:

* This dataset contains information on startup name,industry vertical,location,investors of the startup etc.
* We try to understand the dataset by answering questions like which startup is funded the most,which is the prefered industry verical for investors,which city are startups prefering etc..

Loading the required libraries,data and glimpse into the dataset:

library(plyr)  
library(dplyr)  
library(ggplot2)  
library(ggthemes)  
library(knitr)  
library(kableExtra)  
library(cowplot)  
library(tidyr)  
library(stringr)  
library(zoo)  
library(splitstackshape)  
library(viridis)  
library(ggrepel)  
library(forcats)  
star=read.csv("indian\_startup\_funding (1).csv",header=TRUE,stringsAsFactors = FALSE)  
glimpse(star)

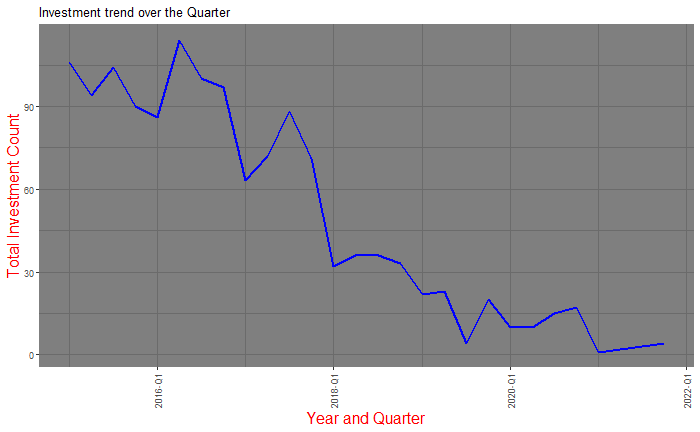
Rows: 3,212  
Columns: 8  
$ Date <chr> "4/14/2021", "4/14/2021", "4/7/2021", "4/5/2021", "4/1/2021", "3/31/2021", "3/30/2021",~  
$ StartupName <chr> "Swiggy", "Beldara", "Groww", "Meesho", "BYJUâ\200\231S", "Uniphore", "BYJUâ\200\231S",~  
$ Industry <chr> "Online Food Delivery", "E-commerce", "FinTech", "E-commerce", "Edu-tech", "Technology"~  
$ Sub.vertical <chr> "Online Food Delivery", "Global B2B marketplace", "Investment platform", "Online resell~  
$ Location <chr> "Bengaluru", "Mumbai", "Bengaluru", "Bengaluru", "Bengaluru", "Palo Alto", "Bengaluru",~  
$ Investors <chr> "Amansa Holdings, Carmignac, Falcon Edge Capital, Goldman Sachs, Think Investments", "H~  
$ InvestmentType <chr> "Series J", "Venture", "Series D", "Series E", "Series F", "Series D", "Series F", "Ser~  
$ AmountinUSD <dbl> 343000000, 7400000, 83000000, 300000000, 460000000, 140000000, 460000000, 8000000, 2750~

summary(star)

Date StartupName Industry Sub.vertical Location   
 Length:3212 Length:3212 Length:3212 Length:3212 Length:3212   
 Class :character Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character Mode :character   
   
   
   
   
 Investors InvestmentType AmountinUSD   
 Length:3212 Length:3212 Min. :1.600e+04   
 Class :character Class :character 1st Qu.:5.000e+05   
 Mode :character Mode :character Median :2.000e+06   
 Mean :2.022e+07   
 3rd Qu.:1.000e+07   
 Max. :3.900e+09   
 NA's :990

Which year has seen more fundings?

#star$Date=as.POSIXct(strptime(star$Date,format="%d/%m/%Y")) # converting into date format   
#star$Month=as.factor(format(star$Date,"%m"))  
#star$Year=as.factor(format(star$Date,"%Y"))  
#star$Monyr=paste(star$Year,star$Month,sep="/")  
#temp= star %>% group\_by(Monyr) %>% summarise(Investments=n()) # count by month,year  
#p=ggplot(temp,aes(Monyr,Investments,group=1))+geom\_line(color="red",size=1)+theme\_dark(base\_size=8)+labs(x="Year and Month of Investment",y="Total Investment Count",title="Investment trend over the years")+theme(axis.title=element\_text(color="red",family="sans",size=12),axis.text.x = element\_text(angle=90,vjust=0.5))# Plot of the year,month trend  
#Quaterly trend:  
star$Qtr=as.yearqtr(star$Date,format="%y-Q%q")  
temp=star %>% group\_by(Qtr) %>% summarise(Investment=n()) %>% drop\_na(Qtr)  
q=ggplot(temp,aes(Qtr,Investment,group=1))+geom\_line(color="blue",size=1)+theme\_dark(base\_size=8)+labs(x="Year and Quarter",y="Total Investment Count",title="Investment trend over the Quarter")+theme(axis.title=element\_text(color="red",family="sans",size=12),axis.text.x = element\_text(angle=90,vjust=0.5))+scale\_x\_yearqtr(format="%Y-Q%q")  
plot\_grid(q,ncol=1)

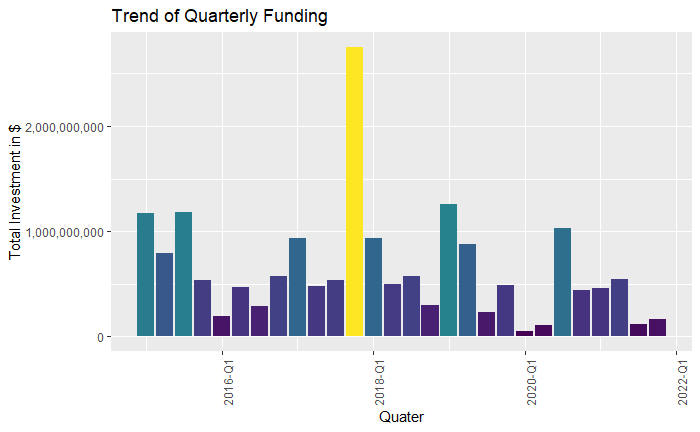


NA  
NA

There is a sharp increase in investment climate after June 2015.Investments have declined sharply during Feb,March of 2017.This is definitely a sharp drop in investments witnessed in Q3 of 2016 and then in 2021.Covid Effect is clearly visible here.

Trend of Quarterly Funding:

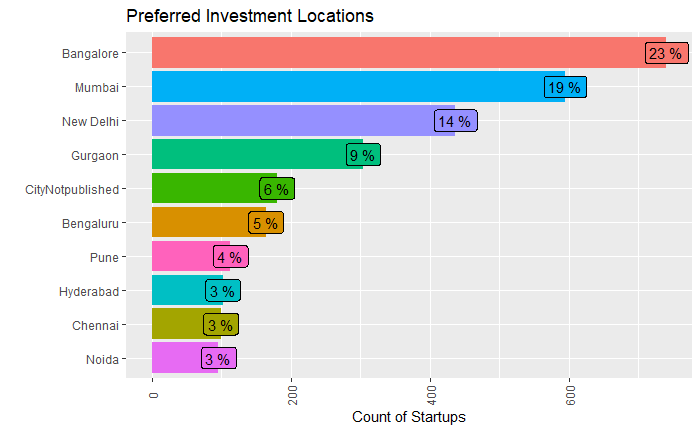
#star$AmountInUSD=as.numeric(gsub(",","",star$Amount In USD)) #Changed from char to numeric and removed comma  
temp=ddply(star,.(Qtr),summarise,TotalInvestments=sum(AmountinUSD,na.rm=TRUE))  
temp%>% drop\_na(Qtr) %>% ggplot(aes(Qtr,TotalInvestments,fill=TotalInvestments))+geom\_bar(stat="identity")+labs(x="Quater",y="Total Investment in $",title="Trend of Quarterly Funding")+theme(axis.text.x=element\_text(angle=90),legend.position="none")+scale\_x\_yearqtr(format="%Y-Q%q")+scale\_fill\_viridis(discrete = FALSE)+scale\_y\_continuous(labels=scales::comma)



The quarter of 2018 was the highest in terms of total investment

Breakdown of Investment Locations:

temp=star %>% select(Location) %>% group\_by(Location) %>% summarise(count=n()) %>% mutate(perc=round((count/sum(count))\*100)) %>% arrange(desc(count)) # Top 10 Cities  
temp[temp==""]="CityNotpublished" # replace blank with this   
temp$perc=paste(temp$perc,"%") # add % to the column  
ggplot(head(temp,10),aes(reorder(Location,count),count,fill=Location))+geom\_bar(stat="identity")+theme(legend.position="none",axis.text.x = element\_text(angle=90,vjust=0.5))+labs(x="",y="Count of Startups",title="Preferred Investment Locations")+geom\_label(aes(label=perc))+coord\_flip() # visualise along with %

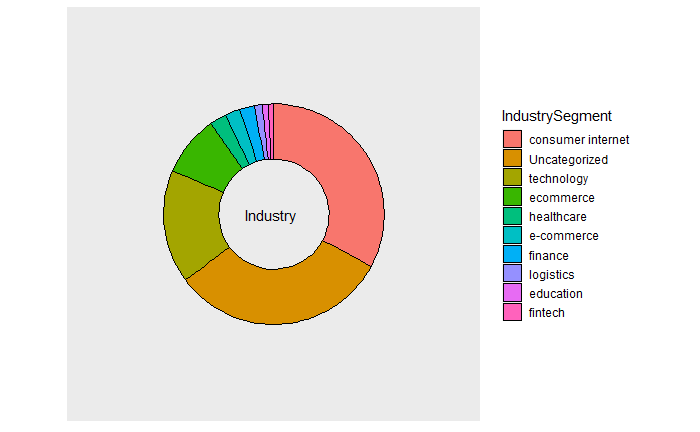


Bangalore, Mumbai, New Delhi had the most startups origins compared to others in India

Breakdown of Industry Vertical:

star$Industry=tolower(star$Industry)  
star$Industry[star$Industry==""]="Uncategorized"  
temp=star %>% select(Industry) %>% group\_by(Industry) %>% drop\_na(Industry)%>% summarise(count=n()) %>% mutate(perc=(count/sum(count))\*100) %>% arrange(desc(count))#finding percentage and count  
temp$ymax=cumsum(temp$perc) #for donut  
temp$ymin=c(0,head(temp$ymax,n=-1))  
temp$Industry=factor(temp$Industry,levels=temp$Industry)  
ggplot(head(temp,10),aes(fill=Industry,ymax=ymax,ymin=ymin,xmax=10,xmin=5))+geom\_rect(color="black")+coord\_polar(theta="y")+xlim(c(0,15))+geom\_label\_repel(aes(label=paste(round(perc),"%"),x=7,y=(ymax+ymin)/2),inherit.aes=TRUE,show.legend=FALSE)+theme(panel.grid = element\_blank(),axis.text = element\_blank(),axis.ticks = element\_blank(),legend.position="right",axis.title=element\_blank()) +annotate("text", x = 0, y = 0 ,label = "Industry ")+labs(fill="IndustrySegment")+guides(fill=guide\_legend(keywidth=1,keyheight=1))

Warning: ggrepel: 2 unlabeled data points (too many overlaps). Consider increasing max.overlaps



Consumer Internet is the prefered model of business for 33 % of the startups followed by technology and e commerce.The data is highely efficient for the analysis of 7 years.

Breakdown of Investment Type:

unique(star$InvestmentType)

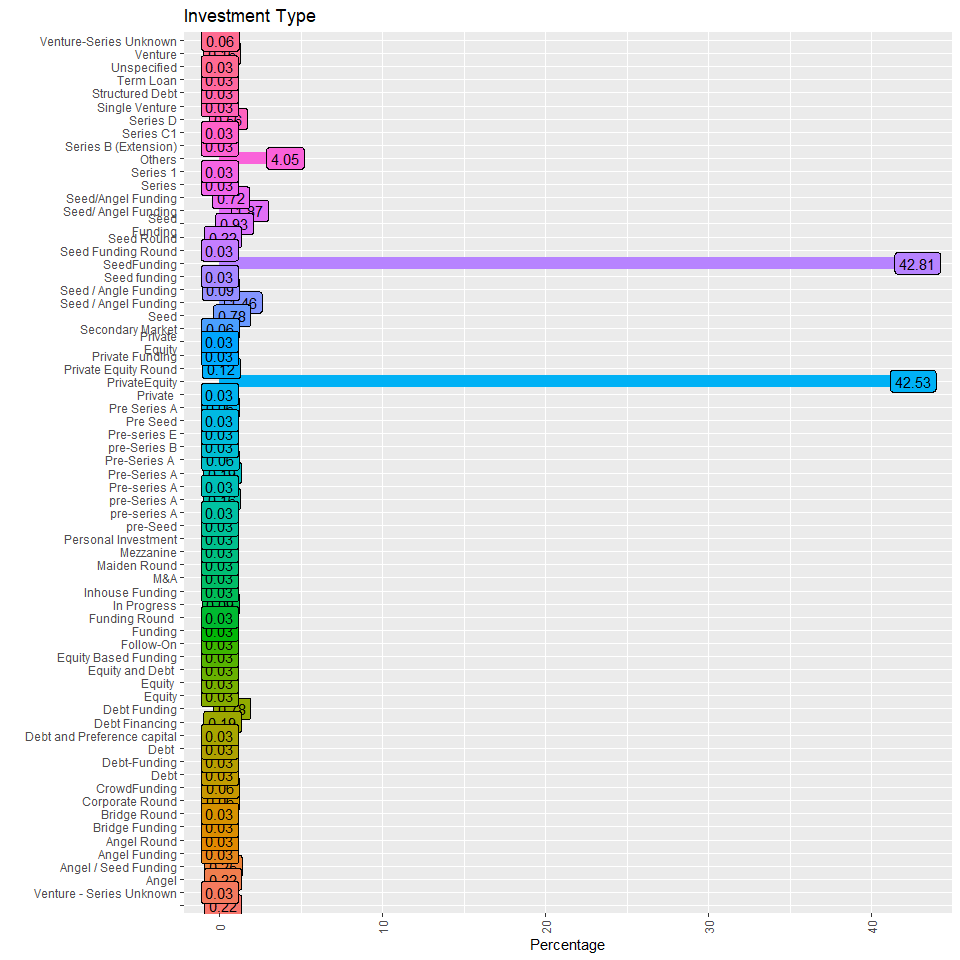
[1] Series J Venture Series D   
 [4] Series E Series F Series A   
 [7] Seed Debt Financing Series C   
[10] Unspecified Series B Venture Round   
[13] SeedFunding Series Secondary Market   
[16] Funding Pre-Series A Bridge Funding   
[19] Pre-series E Venture Equity and Debt   
[22] Personal Investment Follow-On Debt   
[25] Angel pre-Series A   
[28] PrivateEquity In Progress Series D1   
[31] M&A pre-Seed Venture-Series Unknown   
[34] pre-Series B Venture - Series Unknown Pre Seed   
[37] Pre Series A Series C1 Series 1   
[40] Venture Series Venture Series Private Equity Round   
[43] Seed Round Pre-series A Series G   
[46] Debt Funding Funding Round Series H   
[49] Corporate Round Maiden Round pre-series A   
[52] Seed Funding Round Single Venture Angel Round   
[55] Venture - Series Unknown Bridge Round Debt   
[58] Debt and Preference capital Inhouse Funding Seed/ Angel Funding   
[61] Debt-Funding Pre-Series A Equity Based Funding   
[64] Series B (Extension) Mezzanine Equity   
[67] Private Funding Seed / Angel Funding Seed/Angel Funding   
[70] Equity Seed / Angle Funding Seed funding   
[73] Angel / Seed Funding Term Loan Structured Debt   
[76] Private Angel Funding Seed\nFunding   
[79] CrowdFunding Private\nEquity   
80 Levels: Venture - Series Unknown Angel Angel / Seed Funding Angel Funding Angel Round ... Venture Series

There seems to be duplicate values and we need to clean it a bit.For eg.CrowdFunding and Crowd Funding are same but it is interpreted differently.Lets clean it and resummarise.

star$InvestmentType=as.factor(star$InvestmentType)  
star=star %>% mutate(InvestmentType=fct\_collapse(InvestmentType,"PrivateEquity"=c("PrivateEquity","Private Equity"),"SeedFunding"=c("SeedFunding","Seed Funding"),"CrowdFunding"=c("Crowd Funding","Crowd funding"),"Others"=c("Venture Series","Venture Round","Venture - Series Unknown","Venture Series ","Venture ","Series A","Series B","Series C","Series D1","Series E","Series F","Series G","Series H","Series J")))#rename factor

Warning: Problem with `mutate()` column `InvestmentType`.  
i `InvestmentType = fct\_collapse(...)`.  
i Unknown levels in `f`: Private Equity, Seed Funding, Crowd Funding, Crowd funding, Venture Series, Venture Round, Venture - Series Unknown, Venture Series

temp = star %>% group\_by(InvestmentType)%>% summarise(count=n()) %>% mutate(perc=(round(count/sum(count)\*100,2))) %>% arrange(desc(count))  
  
ggplot(temp,aes(InvestmentType,perc,fill=InvestmentType)) +geom\_bar(stat="identity")+theme(legend.position="none",axis.text.x = element\_text(angle=90,vjust=0.5))+labs(x="",y="Percentage",title="Investment Type")+geom\_label(aes(label=perc))+coord\_flip()



NA

More than 42 % of the investors have seed/Angel funded the startups while 42 % of them are through private equity.And 16% are included in others.

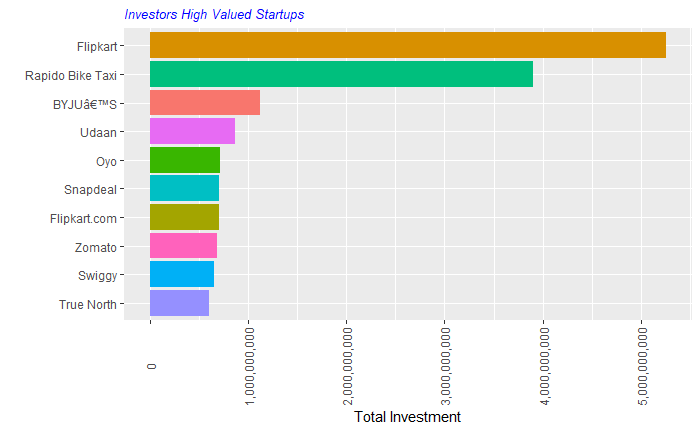
Which startup has investors confidence ?

temp=star %>% select(StartupName,Investors,AmountinUSD) %>% ddply(.(StartupName),summarise,sum=sum(AmountinUSD)) %>% arrange(desc(sum))  
kable(head(temp,25),"html") %>% kable\_styling("striped",full\_width=T) %>% column\_spec(1:2,bold=T,background="white") %>% row\_spec(c(1,2,3,5,6,24,25),bold=F,color="white",background="#ffb6c1")

| StartupName | sum |
| --- | --- |
| **Flipkart** | **5259700000** |
| **Rapido Bike Taxi** | **3900000000** |
| **BYJUâ€™S** | **1120000000** |
| **Ola** | **984500000** |
| **Udaan** | **870000000** |
| **Flipkart.com** | **700000000** |
| **Snapdeal** | **700000000** |
| **Zomato** | **685000000** |
| **Swiggy** | **656500000** |
| **True North** | **600000000** |
| **BigBasket** | **507000000** |
| **Olacabs** | **400000000** |
| **Grofers** | **352000000** |
| **Oyo Rooms** | **350000000** |
| **Automation Anywhere** | **300000000** |
| **SkilloVilla** | **300000000** |
| **FirstCry** | **296000000** |
| **OYO Rooms** | **285000000** |
| **Vogo Automotive** | **283000000** |
| **Edelweiss** | **270000000** |
| **PolicyBazaar** | **267700000** |
| **DealShare** | **250000000** |
| **Delhivery** | **240000000** |
| **Zilingo** | **235900000** |
| **Lenskart.com** | **231000000** |

An birdseye view of the data indicates that the startup variable requires some cleaning.We try to remove .com,in from the variable and combine certain rows since they represent the same company.Flipcart, Udaan, ByJUS, Zilingo, Lenskart.com are booming the market and showing the Investor’s confidence in them.

star$StartupName=as.factor(star$StartupName)  
star =star %>% mutate(StartupName=fct\_collapse(StartupName,"Oyo"=c("Oyo","Oyorooms","Oyo Rooms","OyoRooms"),"Ola"=c("Ola","Olacabs","Ola Cabs")))  
temp=star %>% select(StartupName,Investors,AmountinUSD) %>% ddply(.(StartupName),summarise,sum=sum(AmountinUSD)) %>% arrange(desc(sum))  
#Top 10   
ggplot(head(temp,10),aes(reorder(StartupName,sum),sum,fill=StartupName))+geom\_bar(stat="identity")+theme(legend.position="none",axis.text.x = element\_text(angle=90,vjust=0.5),plot.title=element\_text(color="blue",size=10,face="italic"))+labs(x="",y="Total Investment",title="Investors High Valued Startups")+coord\_flip()+scale\_y\_continuous(labels=scales::comma)



Flipcart is the highest valued Startup followed by Rapido

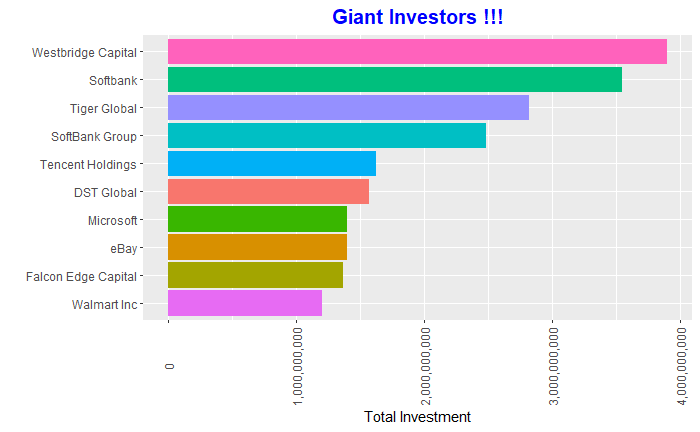
Top Investors:

To find out the top investors,we need to split the investor name column so that each column consists of 1 investor name instead of several investors.The *splitstackshape* package comes in handy for this scenerio since the number of colums to be split is not known because there are unequal number of investors for each startup.We use the cSplit function to split the columns.

star=cSplit(star,"Investors",sep=",",direction="long",drop=FALSE)

Warning in type.convert.default(unlist(x, use.names = FALSE)) :  
 'as.is' should be specified by the caller; using TRUE

temp=ddply(star,.(Investors),summarise,count=n(),TotalInvestment=sum(AmountinUSD))  
temp=temp %>% arrange(desc(TotalInvestment)) %>% drop\_na(TotalInvestment)  
p=ggplot(head(temp,10),aes(reorder(Investors,TotalInvestment),TotalInvestment,fill=Investors))+geom\_bar(stat="identity")+theme(legend.position="none",axis.text.x = element\_text(angle=90,vjust=0.5),plot.title=element\_text(color="blue",size=15,face="bold",hjust=0.5))+labs(x="",y="Total Investment",title="Giant Investors !!!")+coord\_flip()+scale\_y\_continuous(labels=scales::comma)  
plot\_grid(p,align='v',ncol=1)



#Trend of Unique Investors  
#temp=star %>% select(Qtr,Investors) %>% group\_by(Qtr) %>% drop\_na(Qtr)%>% mutate(count=length(unique(Investors)))  
#q=ggplot(temp[(duplicated(c("Qtr","count"))),],aes(Qtr,count))+geom\_line(color="red",size=1)+scale\_x\_yearqtr(format="%Y-Q%q")+theme(legend.position="none",#axis.text.x = element\_text(angle=90,vjust=0.5),plot.title=element\_text(color="blue",size=15,face="bold",hjust=0.5))+labs(x="",y="",title="Trend of Unique #Investors")  
#plot\_grid(q,align='v',ncol=1)

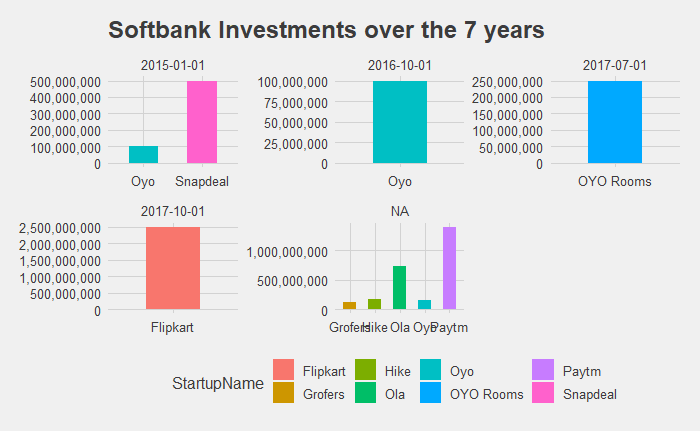
When we the Investors, Westbeidge Capitals is leading the market. While, Softbank and tiger global are also the giant investors. Investor funding took an increase with more than 400+ unique investors entering into startup ecosystem in Q1 of 2016 which has shown a declining trend since then.

We select the top 5 investors to know in which startups they have pumped their investments.

Investor Funding

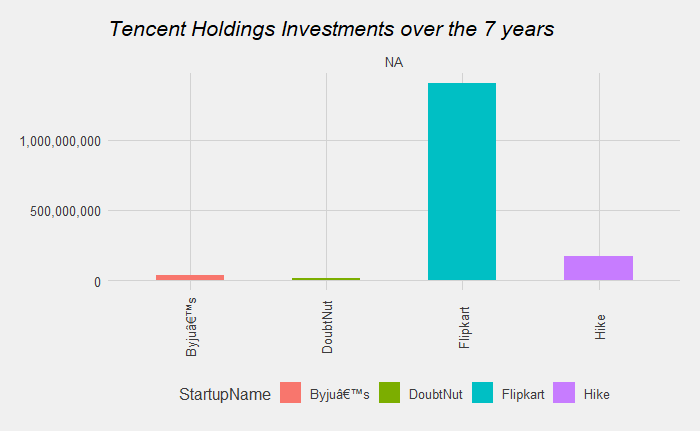
Softbank Group

temp= star %>% filter(Investors=="SoftBank Group" | Investors=="Softbank") %>% select(Qtr,StartupName,AmountinUSD) %>% ddply(.(Qtr,StartupName),summarise,TotalInvestment=sum(AmountinUSD))  
ggplot(temp,aes(x=StartupName,y=TotalInvestment,fill=StartupName))+geom\_bar(stat="identity",position="identity",width=0.5)+facet\_wrap(~as.Date(Qtr),scales=c("free"))+scale\_y\_continuous(labels=scales::comma)+theme(legend.position="none",axis.text.x = element\_text(angle=90,vjust=0.5),plot.title=element\_text(color="blue",size=10,face="italic"))+labs(x="",y="Total Investment",title="Softbank Investments over the 7 years")+theme\_fivethirtyeight()



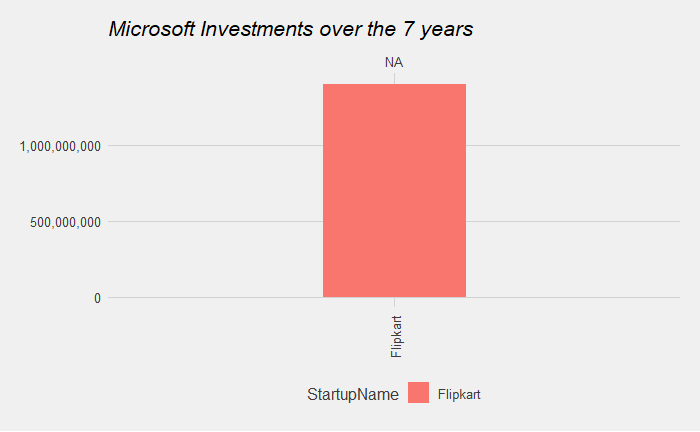
Tencent Holdings

temp= star %>% filter(Investors=="Tencent Holdings") %>% select(Qtr,StartupName,AmountinUSD) %>% ddply(.(Qtr,StartupName),summarise,TotalInvestment=sum(AmountinUSD))  
ggplot(temp,aes(x=StartupName,y=TotalInvestment,fill=StartupName))+geom\_bar(stat="identity",position="identity",width=0.5)+facet\_wrap(~as.Date(Qtr),scales=c("free"))+scale\_y\_continuous(labels=scales::comma)+theme\_fivethirtyeight()+theme(axis.text.x = element\_text(angle=90,vjust=0.5),plot.title=element\_text(color="black",size=16,face="italic"))+labs(x="",y="Total Investment",title="Tencent Holdings Investments over the 7 years")



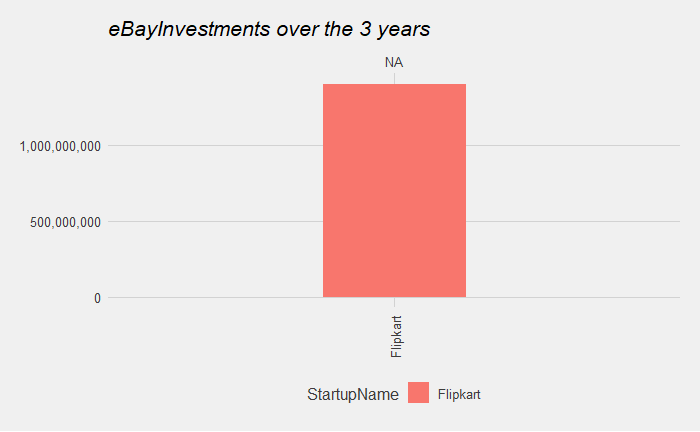
Microsoft

temp= star %>% filter(Investors=="Microsoft") %>% select(Qtr,StartupName,AmountinUSD) %>% ddply(.(Qtr,StartupName),summarise,TotalInvestment=sum(AmountinUSD))  
ggplot(temp,aes(x=StartupName,y=TotalInvestment,fill=StartupName))+geom\_bar(stat="identity",position="identity",width=0.3)+facet\_wrap(~as.Date(Qtr),scales=c("free"))+scale\_y\_continuous(labels=scales::comma)+theme\_fivethirtyeight()+theme(axis.text.x = element\_text(angle=90,vjust=0.5),plot.title=element\_text(color="Black",size=16,face="italic"))+labs(x="",y="Total Investment",title="Microsoft Investments over the 7 years")



eBay

temp= star %>% filter(Investors=="eBay") %>% select(Qtr,StartupName,AmountinUSD) %>% ddply(.(Qtr,StartupName),summarise,TotalInvestment=sum(AmountinUSD))  
ggplot(temp,aes(x=StartupName,y=TotalInvestment,fill=StartupName))+geom\_bar(stat="identity",position="identity",width=0.3)+facet\_wrap(~as.Date(Qtr),scales=c("free"))+scale\_y\_continuous(labels=scales::comma)+theme\_fivethirtyeight()+theme(axis.text.x = element\_text(angle=90,vjust=0.5),plot.title=element\_text(color="black",size=16,face="italic"))+labs(x="",y="Total Investment",title="eBayInvestments over the 3 years")



Conclusion

The year 2015 seems to be a boom period for startup investments but this has seen a downward trend during the entire 2016.Lot of factors can be attributed to this.

One event which shook up the Indian Ecosystem is the Demonitisation move by the government in Nov 2016.

This project has focused to analyse the dataset from differnt angles and has partly tried to find out the top players in the startup space as far as the investments are concerned.

As the quarter of 2018 was the highest in terms of total investment, but covis made that breakthrough. Though Investors showed interests, market was still recovering in 2020.

The startups in 2021 seems to be stabled.

**Thank you for reading our Analysis on Indian Startup Funding**

