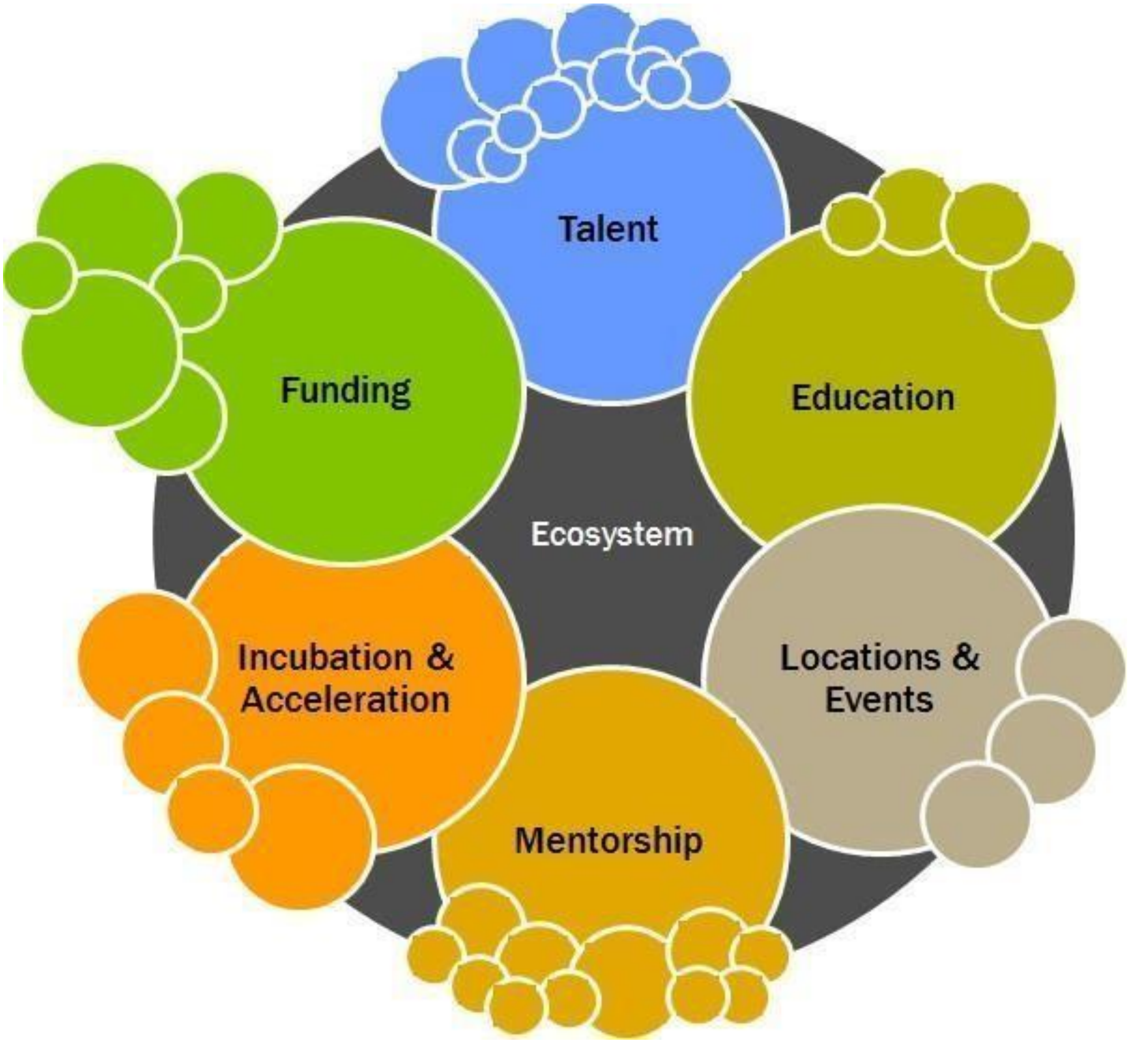
A project report by Ashish S. Parpolkar

INDEX

Indian
START UP
Ecosystem


Talent
Funding
Education
Ecosystem
Incubation & Acceleration
Locations & Events
Mentorship

INTRODUCTION

A startup is a newly established company whose corporate form allows clear distinction between different shareholders and the ability to receive outside financing while the goal is super rapid growth and market dominance by offering an innovative product or service. The term startup refers to a company in the first stages of operations. Startups are founded by one or more entrepreneurs who wantto develop a product or service for which they believe there is demand. These companies generally start with high costs and limited revenue, which is why they look for capital from a variety of sources such as venture capitalists. India is the world's sixth-largest economy by GDP after the United States, China, Japan, Germany, and France. India has the 3rd largest startup ecosystem in the world.

India has about 50,000 startups in India in 2018, around 8,900 – 9,300 of these are technology led startups. 1300 new tech startups were born in 2019 alone implying there are 2-3 tech startups born every day. Bangalore has emerged as India's primary startup hub, but significant founding activity is also taking place inMumbai and the National Capital Region (NCR), as well as some smaller cities.

Further, the study investigates how the startup ecosystem has developed over theyears and describes where and which kind of support is available.

India has emerged as a major player in the global start-up scene over the past decade. The country's favorable regulatory environment, access to capital, anda large pool of talented entrepreneurs have contributed to the growth of the start- up ecosystem. According to a report by NASSCOM, India is the third-largest start-up ecosystem in the world, with over 50,000 start-ups in operation.

The Indian start-up funding ecosystem has also grown significantly in recent years, with a significant increase in the number of angel investors, venture capitalists, and private equity firms investing in the space. In 2021, Indian start- ups raised a total of $14.5 billion in funding across 681 deals, according to data from Venture Intelligence. This represents a substantial increase from 2020, whenIndian start-ups raised $11.7 billion in funding across 1,021 deals.

The sectors that attracted the most funding in 2021 were e-commerce, fintech, edtech, and health tech. Flipkart, Paytm, BYJU'S, and Pharm Easy were among the top-funded start-ups in these sectors. However, there has also been arise in funding for start-ups in sectors such as agritech, cleantech, and deep tech,which are poised to make a significant impact on the Indian economy.

The Indian government has played a crucial role in fostering the start-upecosystem, with initiatives such as Startup India, which provides tax benefits andfunding to start-ups, and the creation of a Rs.10,000 crore ($1.3 billion) fund of funds to provide early-stage funding to start-ups. The government has also launched several other schemes, such as the Atal Innovation Mission and DigitalIndia, which have helped to create a supportive environment for start-ups.

Despite the positive developments, the Indian start-up ecosystem still faces several challenges. One of the primary challenges is a lack of access to early-stage funding, which is essential for start-ups to grow and scale. Another significant challenge is regulatory barriers, which can impede the growth of start- ups. For example, India's complex tax laws and bureaucratic procedures can make it difficult for start-ups to operate efficiently. Additionally, there is a shortage of skilled talent, particularly in emerging areas such as artificial intelligence and block chain.

However, with the continued growth of the ecosystem and supportive government policies, India's start-up funding ecosystem is poised for further growth and success in the coming years. The rise of start-ups in sectors such as agritech, cleantech, and deep tech presents a tremendous opportunity for India tobecome a global leader in these areas. With the right policies and support, the Indian start-up ecosystem can continue to thrive and contribute to the country's economic growth and development.

OBJECTIVES

- How does the funding ecosystem change with time.
- Do cities play an important role in funding.
- Which industries are favored by investors for funding.
- Who are the important investors in the Indian Ecosystem.
- To fit ARIMA model to the Funding data.
- To predict future growth in industry sector by using ARIMA model.

RESEARCH

1. Data Source:

To understand and analysis the start-up funding ecosystem in India, data which is secondary in nature have been collected from various websites duringthe period of March 2015-April 2021

https://trak.in/india-startup-funding-investment-2015/

2. VARIABLE DESCRIPTION :

A. Startup Name: Name of the Company

B. Industry Vertical: An industry vertical, however, is more specific and describes a group of companies that focus on a shared niche or specialized market spanning multiple industries. Also called vertical markets.

C. City Location: City in which startup was started.

D. Investors Name: Name of the investors who invested in startups.

E. Investment Type: which type of investment made by investors. There are types of investment like Series A, B, C, D, E, F, . . ., seed funding, debt funding, Venture etc.

F. Amount: Invested amount by investors in Crore.

METHODOLOGY

I. Introduction to Time Series Forecasting

A time series is a sequence where a metric is recorded over regular time intervals. Depending on the frequency, a time series can be of yearly, quarterly, monthly, weekly, daily, hourly, minutes and even seconds wise.

Forecasting is a time series often of tremendous commercial value. In most manufacturing companies, it drives the fundamental business planning, procurement and production activities. Any errors in the forecasts will ripple down throughout the supply chain or any business context for that matter. So it's important to get the forecasts accurate in order to save on costs and is critical to success. Not just in manufacturing, the techniques and concepts behind time series forecasting are applicable in any business.

Now forecasting a time series can be broadly divided into two types.

If you use only the previous values of the time series to predict its future values, it is called **Univariate Time Series Forecasting**. And if you use predictors other than the series (a.k.a. exogenous variables) to forecast it is called **Multi Variate Time Series Forecasting**.

This post focuses on a particular type of forecasting method called **ARIMA** modelling. ARIMA, short for 'Auto Regressive Integrated Moving Average', is a forecasting algorithm based on the idea that the information in the past values of the time series can alonebe used to predict the future values.

II. Introduction to ARIMA Models

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values. Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.

An ARIMA model is characterized by 3 terms :p, q, d

Where,
p is the order of the AR term
q is the order of the MA term

d is the number of differencing required to make the time series stationary

If a time series, has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for ‘Seasonal ARIMA’. More on that once we finish ARIMA. So, what does the ‘order of AR term’ even mean? Before we go there, let’s first look at the ‘d’ term.

III. What does the p, d and q in ARIMA model mean

The first step to build an ARIMA model is to make the time series stationary.

Because, term ‘Auto Regressive’ in ARIMA means it is a linear regression model that uses its own lags as predictors. Linear regression models, as you know, work best when the predictors are not correlated and are independent of each other.

So how to make a series stationary?

The most common approach is to difference it. That is, subtract the previous value from the current value. Sometimes, depending on the complexity of the series, more than one differencing may be needed. The value of d, therefore, is the minimum number of differencing needed to make the series stationary. And if the time series is already stationary, then d = 0.

Next, what are the ‘p’ and ‘q’ terms

‘p’ is the order of the ‘Auto Regressive’ (AR) term. It refers to the number of lags of Y to be used as predictors. And ‘q’ is the order of the ‘Moving Average’ (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model.

A pure Auto Regressive (AR only) model is one where Yt depends only on its own lags. That is, Yt is a function of the ‘lags of Yt

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

where, Yt-$_1$ is the lag1 of the series, ß is the coefficient of lag1 that the model estimates and α is the intercept term, also estimated by the model. Likewise a pure Moving Average (MA only) model is one where Yt depends only on the lagged forecast errors.

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

where the error terms are the errors of the autoregressive models of the respective lags. The errors Et and E(t-1) are the errors from the following equations:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_0 Y_0 + \epsilon_t$$

$$Y_{t-1} = \beta_1 Y_{t-2} + \beta_2 Y_{t-3} + .. + \beta_0 Y_0 + \epsilon_{t-1}$$

where the error terms are the errors of the autoregressive models of the respective lags. The errors Et and E(t-1) are the errors from the following eq That was AR and MA models respectively.

An ARIMA model is one where the time series was differenced at least once to make itstationary and you combine the AR and the MA terms. So the equation becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

IV. ARIMA model in words:

Predicted Yt = Constant + Linear combination Lags of Y (up to p lags) + Linear Combination of Lagged forecast errors (up to q lags)

The objective, therefore, is to identify the values of p, d and q. But how? Let’s start with finding the ‘d’.

How to find the order of differencing (d) in ARIMA model:

The purpose of differencing it to make the time series stationary.

But you need to be careful to not over-difference the series. Because, an over differenced series may still be stationary, which in turn will affect the model parameters.

So how to determine the right order of differencing?

The right order of differencing is the minimum differencing required to get a near-stationary series which roams around a defined mean and the ACF plot reaches to zero fairly quick. If the autocorrelations are positive for many numbers of lags (10 or more), then the series needs further differencing. On the other hand, if the lag 1 autocorrelation itself is too negative, then the series is probably over-differenced. In the event, you can’t really decide between two orders of differencing, then go with the order that gives the least standard deviation in the differenced series.

STATISTICAL TOOLS AND TECHNIQUES

- **Techniques:**

    1. SHAPIRO–WILK TEST
    2. Time Series Analysis
    3. Decomposition of Series
    4. Tests for stationarity
    5. Identification of ARIMA (p, q) components
    6. Time series model
        a. ARIMA
        b. SRIMA
    7. Model Validation
    8. Forecasting

- **Tools:**

    1. Jupyter notebook
    2. R- programming
    3. Power-BI

5. EXPLORATORY DATA ANALYSIS

1.1 To Check Normal Data:

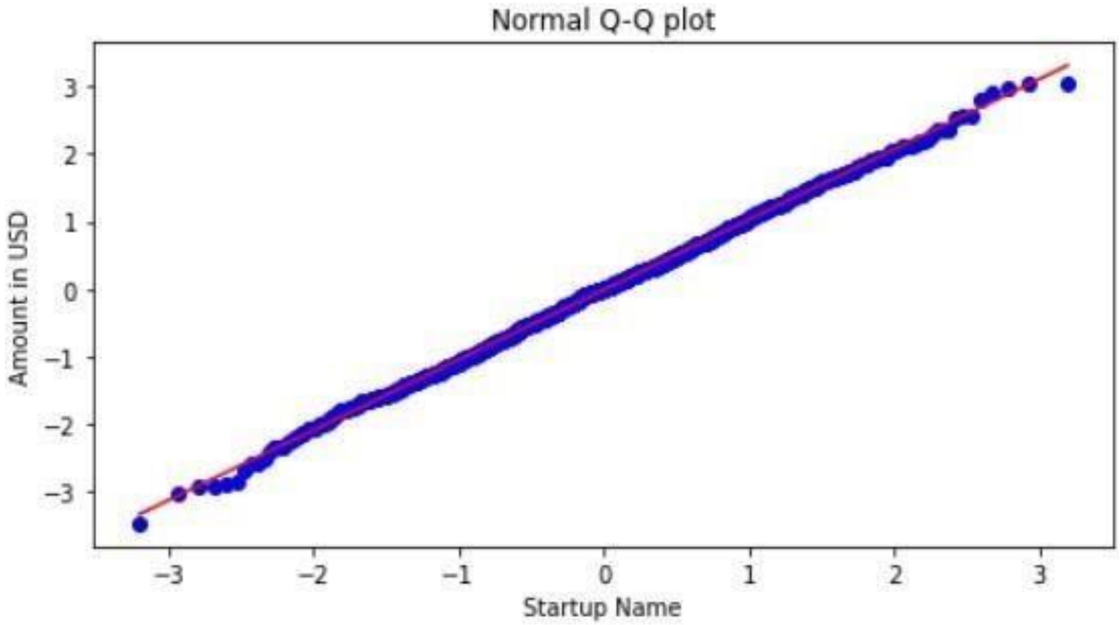


Fig: A

If the Q-Q plot shows a linear pattern, it indicates that the data is normally distributed. This means that the majority of startups receive funding amount that fall within a certain range.

1.2 Top 10 Startups on the basis of funds acquired

Observation Table: -

	Startup Name	Amount In USD
0	Paytm	2340000000
1	Flipkart	2259700000
2	ola	1899500000
3	Snapdeal	700000000
4	Oyo rooms	375000000
5	quirk	230000000
6	delivery	215000000
7	Food panda	210000000
8	shop clues	207700000
9	big basket	207000000

Graph:

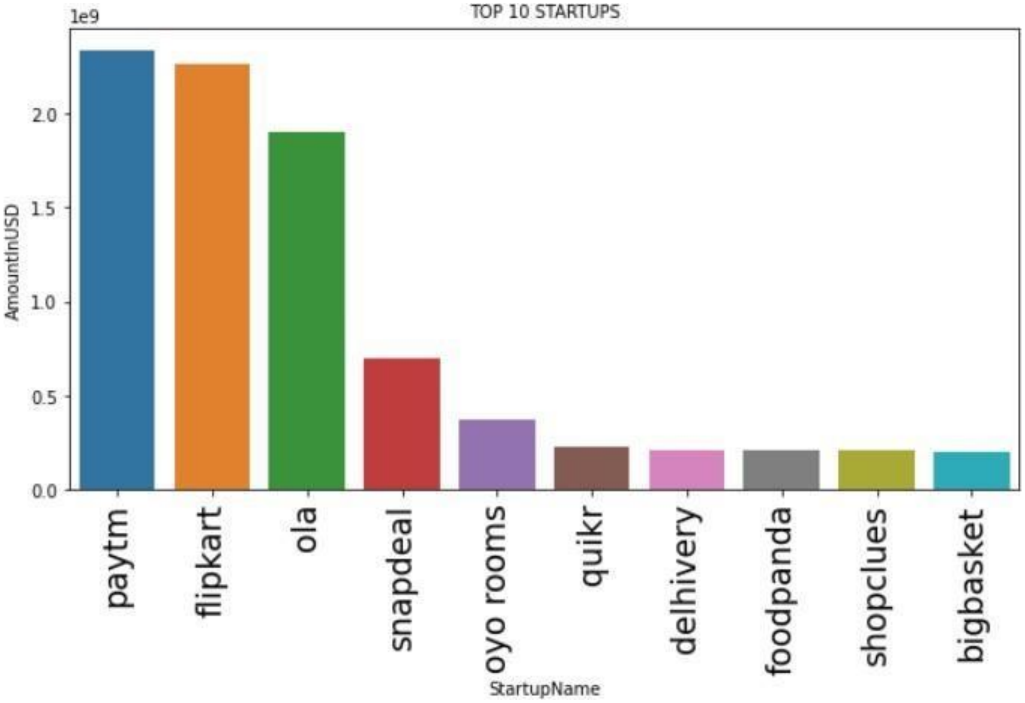


Fig: B

Conclusion:

The above figure A shows the top 10 startup funded between January 2015 to December 2021. Then the Paytm and Flipkart are the highest funded startup.

1.3 Cities Preferred by Investors:

Observation Table:

	City	Number of Startups
1	Bangalore	634
2	Mumbai	449
3	New Delhi	385
4	Gurgaon	241
5	Pune	91
6	Noida	79
7	Hyderabad	77
8	Chennai	67
9	Ahmedabad	35
10	Jaipur	25

Graph: -

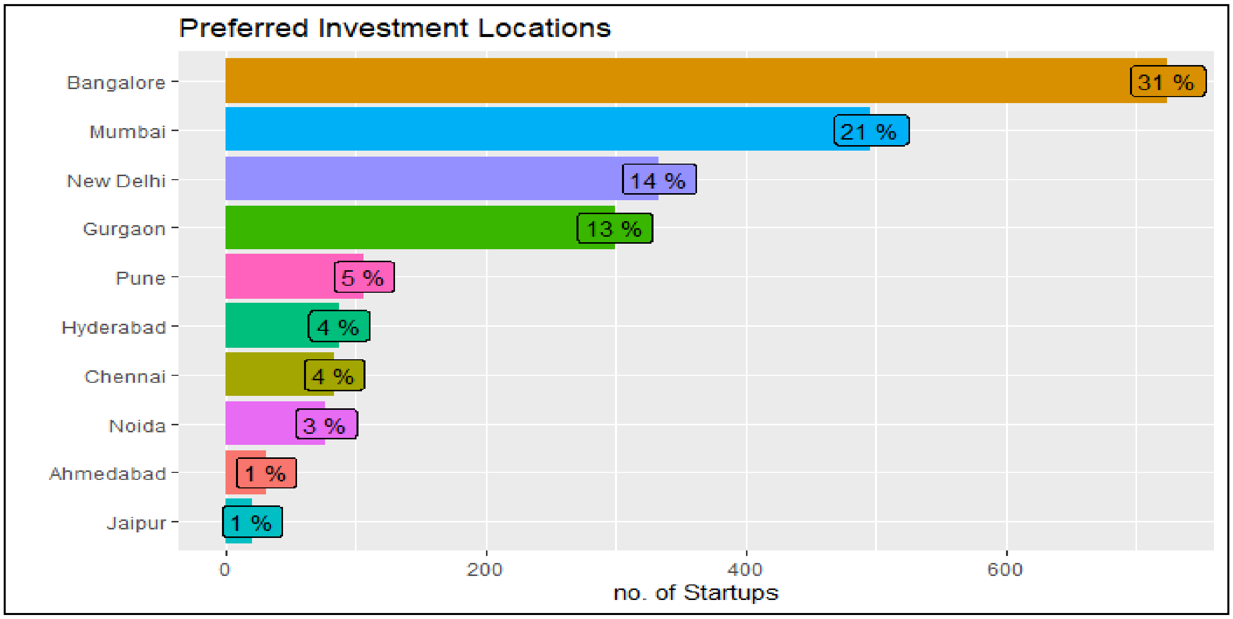


Fig: C

Conclusion: -

Most of the Startups are located in metropolitan cities like Bangalore, Mumbai, New Delhi, because these are the one of the famous Global Startup hubs also Gurgaon, Pune, Hyderabad, Chennai comes under the list of emerging startups. So, the investors preferthese locations for starting their company.

1.4 Industries favored by Investors for funding:

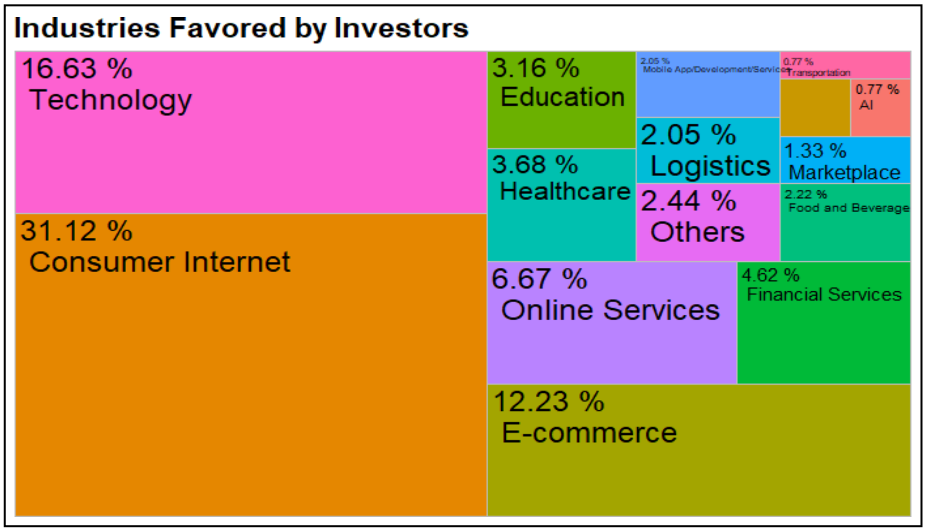


Fig: D

We observe that Consumer Internet related Industries are mostly favored by Investor's. Followed by Technology, E-commerce, Online services, etc. In the above plot we see Education and Healthcare sector is also there says that these are upcoming investment sectors favored by Investor's.

1.5 Industries with high average investment:

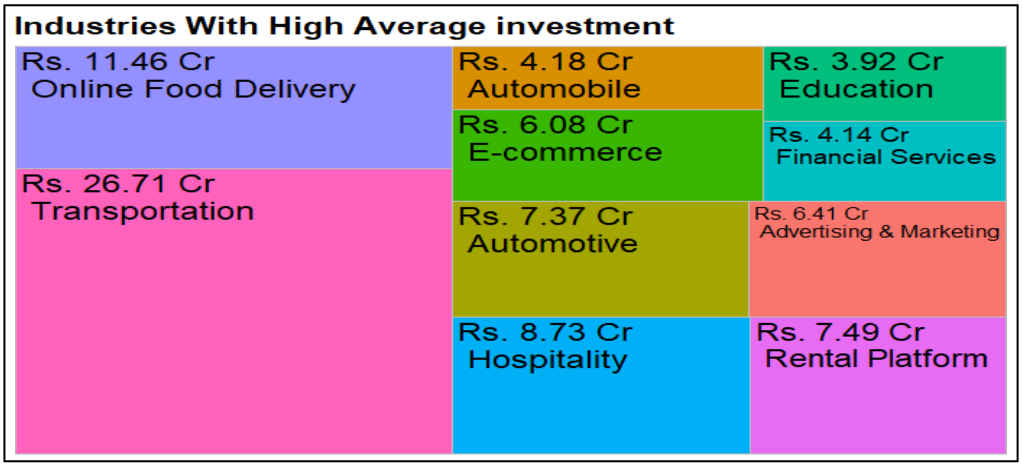


Fig: E

We observe from the above plot of "Industries with High Average Investment" that Transportation, Online Food Delivery has high average investments compare to other. But, Sectors like Hospitality, Rental Platform, Automotive, Advertising and Marketing Also shows high average investment. Education, E-Commerce and Software can be seen as emerging industries which are showing high average investments.

1.6 Breakdown of Investment Type:

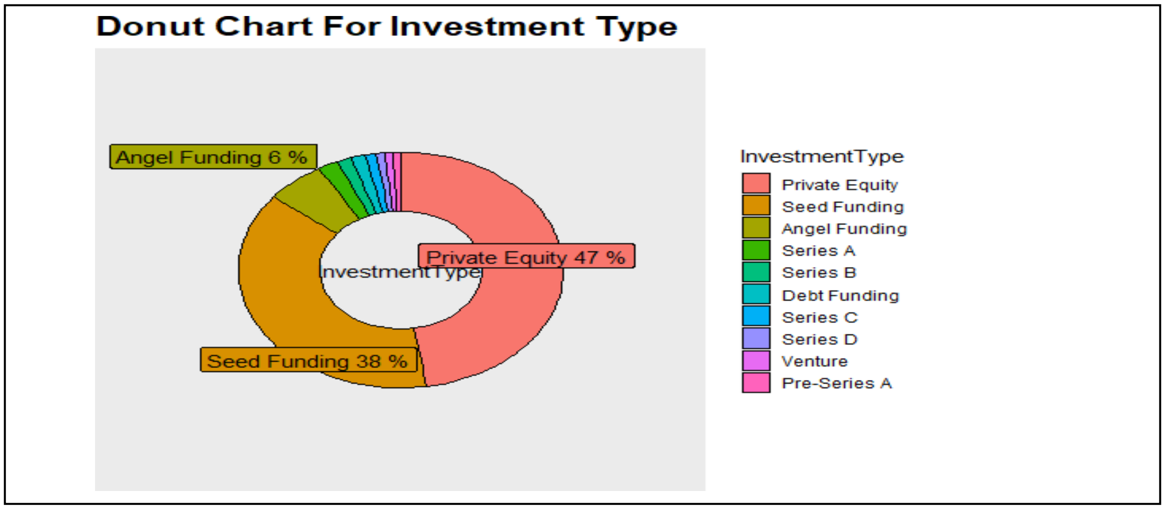


Fig: F

We observe that 47% of Funds are provided by Private Equity, Seed Funding and Angel Funding. Private Equity is favored by companies because it allows them access to liquidity as an alternative to conventional financial mechanisms, such as high interest bank loans or listing on public markets also it has Management incentives, Proven returns, Commitment to success. Followed by Private Equity, Seed funding has more contribution. The most significant advantage of seed capital is that the investors are ready to take the high risk of failure involved in the startup business that shows Investor has faith in the startup idea. Angel funding is there because these fund providers generally want ownership equity in the company in exchange of funds.

1.7 Which Start-up has more investors' confidence:

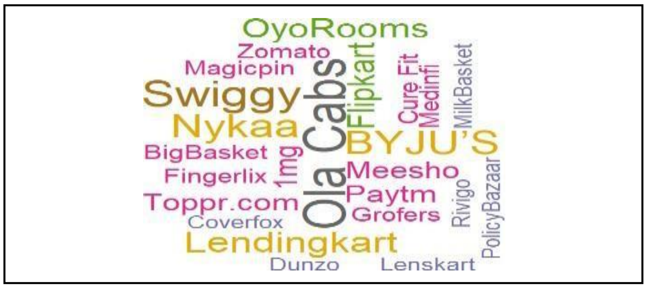


Fig: G

We see that from word cloud that Ola Cabs has more investor's confidence cause more no. of Investors has invested in it followed by Swiggy, BYJU's, Nayaka and so on. i.e. We see that Ola Cabs, Swiggy which are transportation and food delivery servicesectors has more confidence of investors.

2.STATISTICAL ANALYSIS

- **SHAPIRO–WILK TEST:**

We Use Shapiro Wilk Test for Checking Normality of Amount
Null and Alternative hypothesis:
Ho: Amount is normally distributed.
H_1: Amount is Not normally distributed.

Test Statistic is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

We use alpha=5%,

data: D$ Amount

W = 0.16532,

p-value < 2.2e-16

here, P-value is less than alpha=0.05, we reject $H_{0.}$

we conclude that Amount is Not normally distributed.

2.1. Feature Selection

For getting an a more interpretable model it is better to have a selected important variable in the model. That’s why we took a help of feature selection variable selection methods enables us to keep significant variable in model and remove insignificant ones. we applied Kruskal Wallis test with respect to amount

Non-Parametric Tests

- significance of city location on funding amount
- significance of industry vertical on funding amount
- significance of investment amount

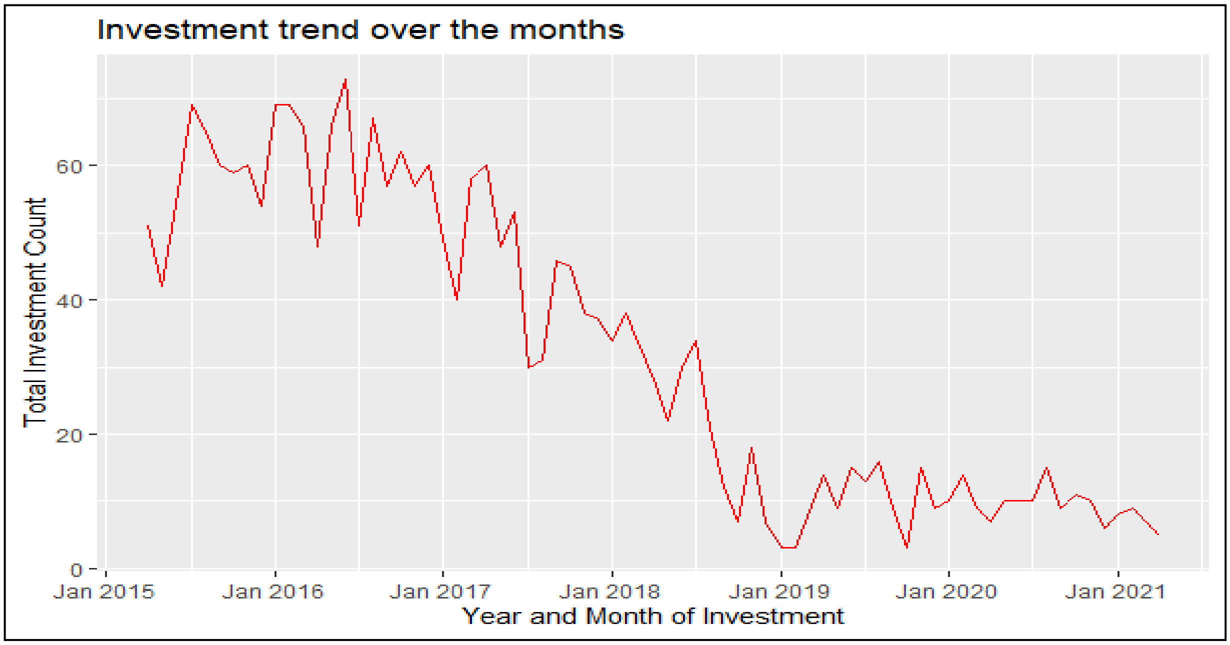
p-values for the test given below 0 Since P-value (=0.14) for

Industry vertical	Sub vertical	City Location	Investors name	Investment type
3.91*10-16	1.42*10-1	1.39*10-7	5.36*10-6	0

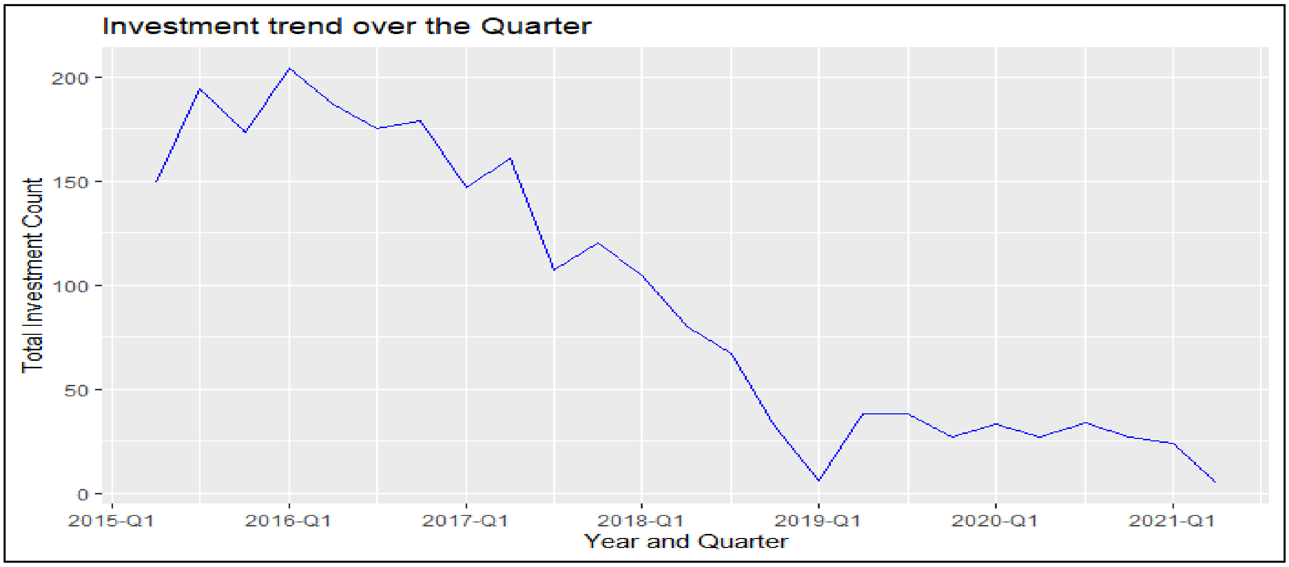
Since P-value =0.14 for subvertical is greater than the level of significance 0.05, therefore subvertical is insignificant with respect to the amount, we can remove it. All other are significant.

2.2 Time Series Analysis

2.2.1 Trend for number of investments over the year:

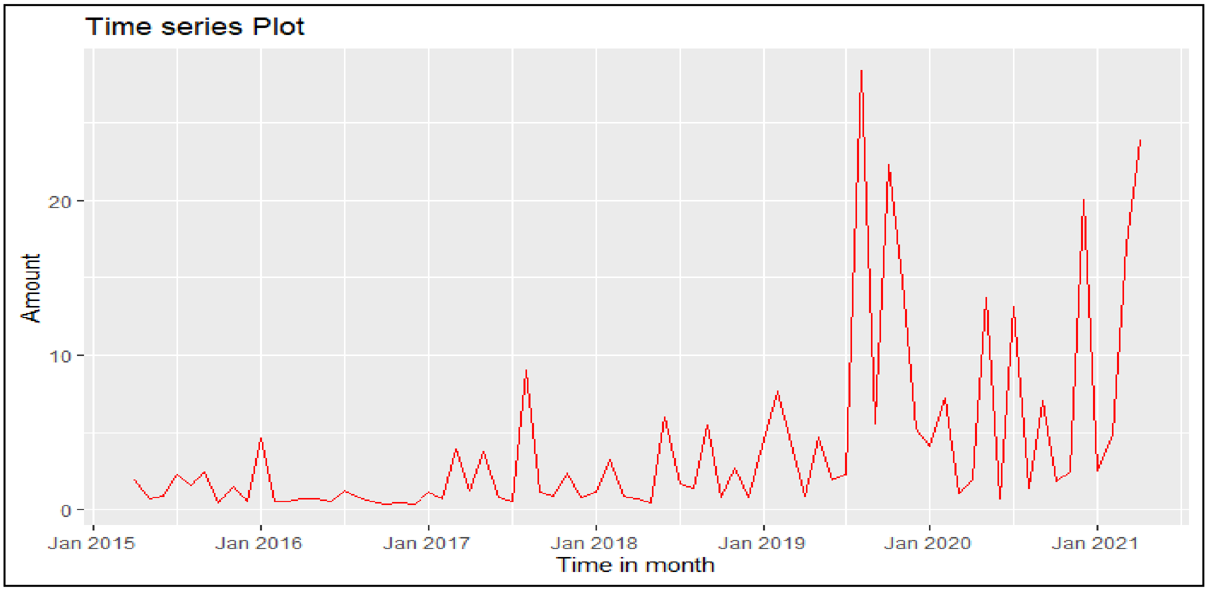


we can clearly see here a decreasing trend in the series but amount of investment is increasing.



Here we can see the most no of investments are in the first quarter of 2016 and the least no of investments are in first quarter of 2019.

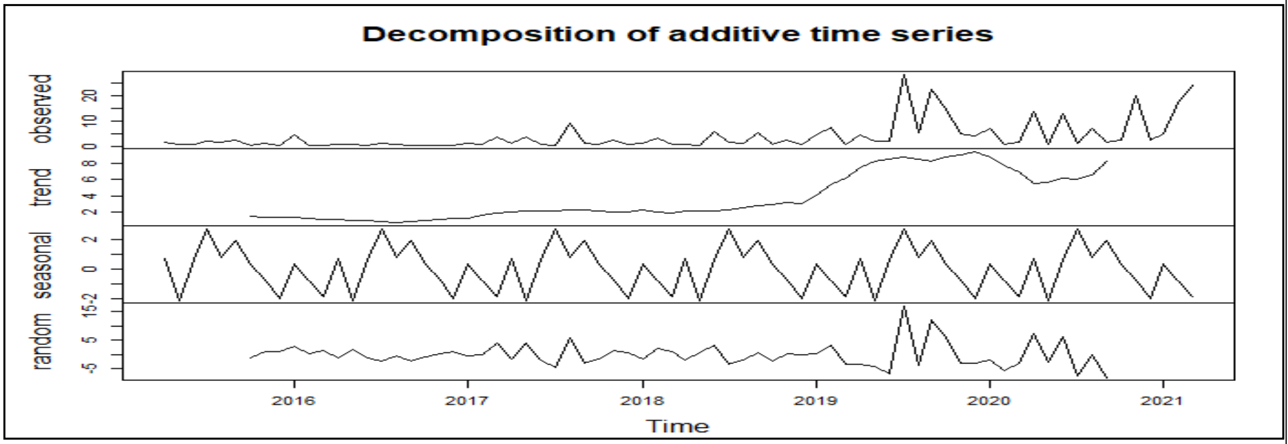
2.2.2 Monthly Total Investment Trend:



We found out one interesting fact, Though the no. Of investment decrease year by year, total amount of investment increases. This happened because of our funding amount is highly depends on investment type (we will show this in below analysis) and hence the startups who already got seed funding or angle funding are now in the position of getting series j or series H type funding and it is obvious that series j or H funding is huge than seedor angel funding.

It is obvious during the pandemic, no. Of investment are likely to be lower. Because no one wants to take a risk in investing onto newly started startup. However, the start up like byjus , ola cab, swiggy who are already got investors' confidence and they are getting series Hor J funding. Therefore, the trends are conversing each other.

2.2.3 Decomposition of Series:



2.2.4 Tests for stationarity:

To test the stationarity of data we use Augmented Dickey Fuller test (adf test) aswell as Phillips and Perron unit root test (PP test). No unit root implies that data is stationary.

Null hypothesis for adf test

H_0: Data is non-stationary

H_1: Data is stationary

Augmented Dickey-Fuller Test

data: series

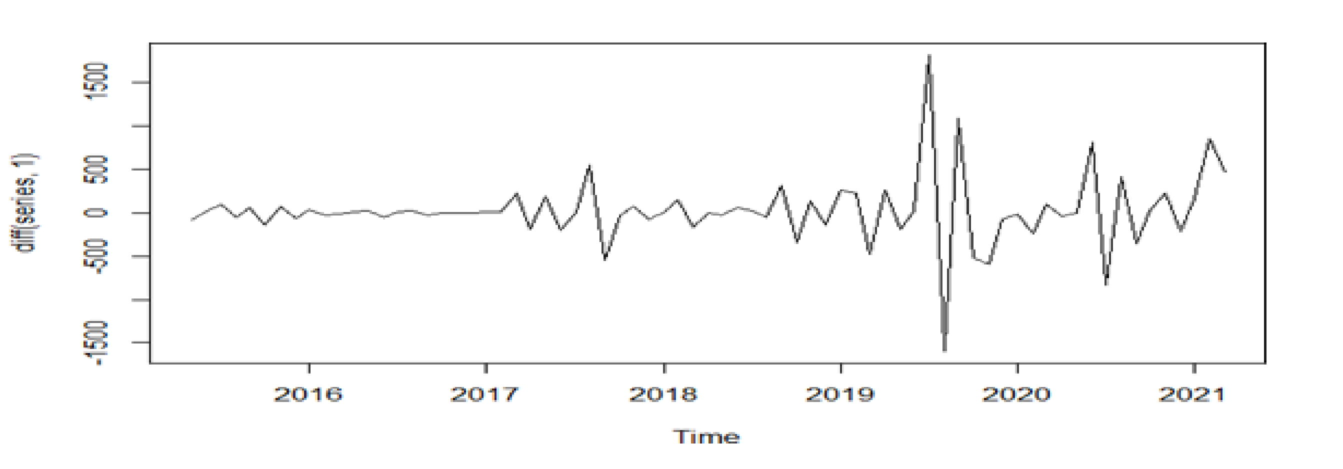
Dickey-Fuller= -3.3255, Lag order=4, p-value=0.01alternative hypothesis: stationary

Phillips-Perron Unit Root Test

data: series
Dickey-Fuller = -6.1775, Truncation lag parameter = 3, p-value = 0.01

From both the test, we conclude that our data is non-stationary. We need to bring stationarity and achieve stationarity we go for differencing.

2.2.5 Identification of level of Differencing:

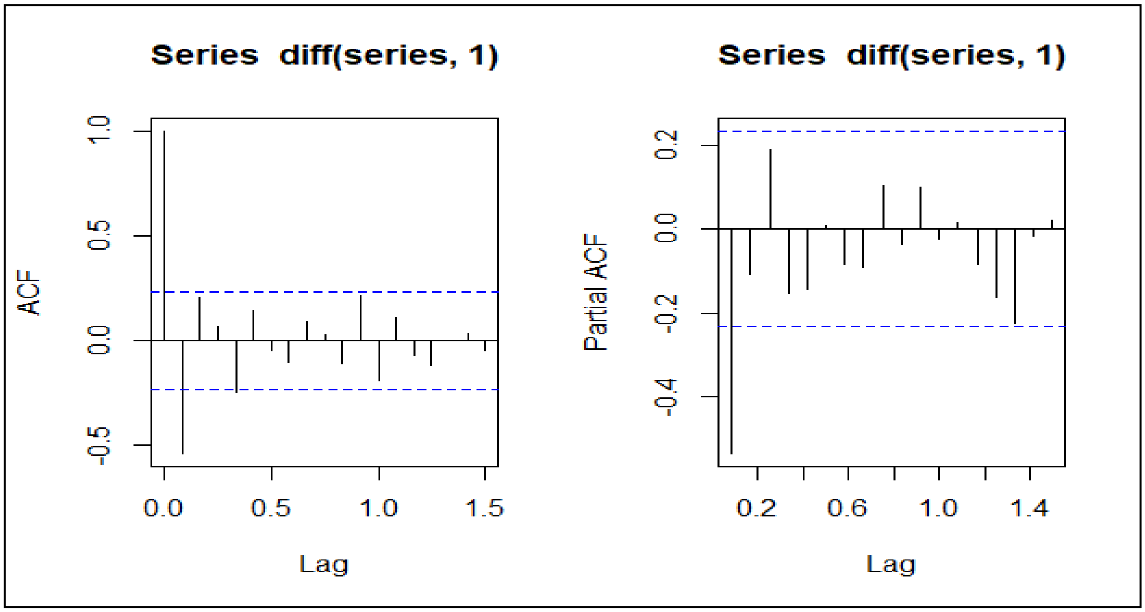


We tried for d=1 checking stationarity for differenced series

Augmented Dickey-Fuller Test

data: series Dickey-Fuller= -3.3255, Lag order=4, p-value=0.0124alternative hypothesis: stationary
adf test for lag one difference shows series is stationary. so, lag one difference in enoughto bring stationarity.

2.2.6 Identification of ARIMA (p, q) components



Basically, we use ACF to identify the MA parameters whereas PACF plot uses to identify the AR parameters. Here we can take MA as 1 or we can take 4 as well because lag 2 and lag 4 auto correlation on the cutoff of significance line. And we can take AR parameter is 1.

2.2.7 Time Series Models:

Model without considering seasonality - ARIMA

(1,1,1)Formula:

ARIMA (Formula:):

$$(1 - B)d\varphi(B)X(t) = \theta(B)Z(t)$$

Where, $\theta(B) = 1 + \theta 1B + \cdot\cdot\cdot + \theta qBq$

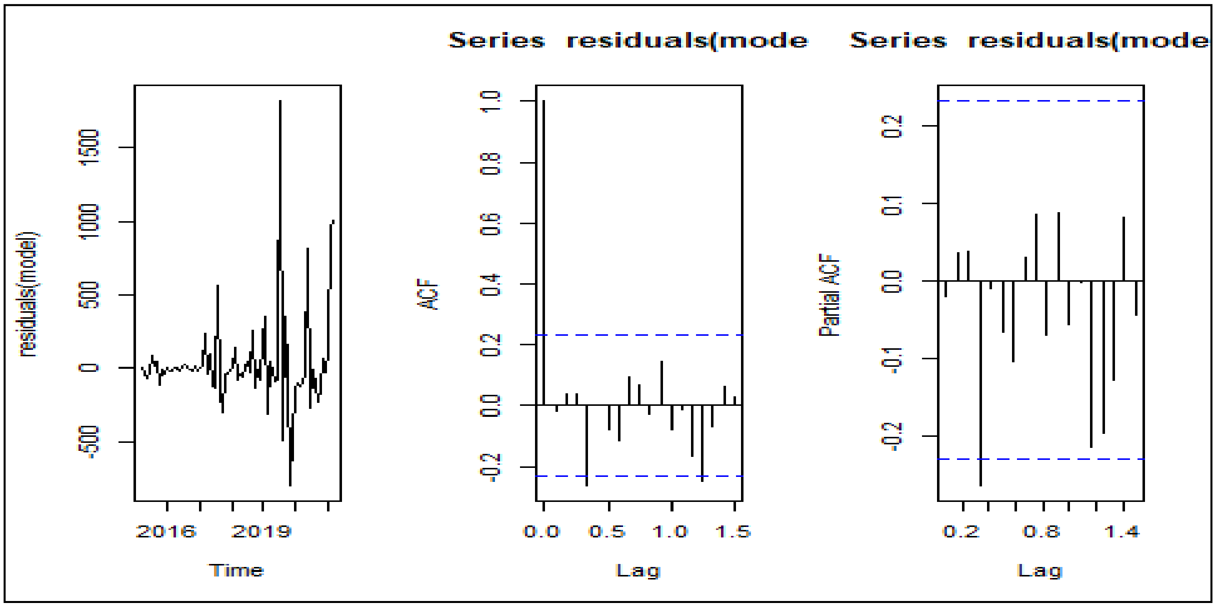
$$\varphi(B) = 1 - \varphi 1B - \cdot\cdot\cdot - \varphi pBp$$

$$(1 - B)Xt = Xt - Xt{-}1$$

Coefficients

Coefficients	ar1	ma1
	-0.3969	-0.2028
S. E	0.1962	0.2196

sigma square estimated as 24.24, log likelihood = -214, AIC =434.22



- Box-Ljung test

  X-squared = 0.15902 d.f = 1, p-value = 0.6901

Above residual plot shows residuals are not uncorrelated. so, we need to gofor model improvement. Let us try to fit SARIMA models.

- SRIMA(1, 1, 1) × (0, 1, 0)12

  Formula:

  SRIMA(p, d, q) × (P, D, Q)S :

  $\Phi(B^S)(1 - B^S)D\varphi(B)(1 - B)dXt = \Theta(B^S)\theta(B)Zt$

  where,

  $\Theta(B^S) = 1 + \Theta_1 B^S + \Theta_2 B^{2S} + \cdot\cdot\cdot + \Theta_Q B^{QS}$

  $\theta(B) = 1 + \theta_1 B + \cdot\cdot\cdot + \theta q B^q$

  $\Phi(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdot\cdot\cdot - \Phi_P B^{P\,S}$

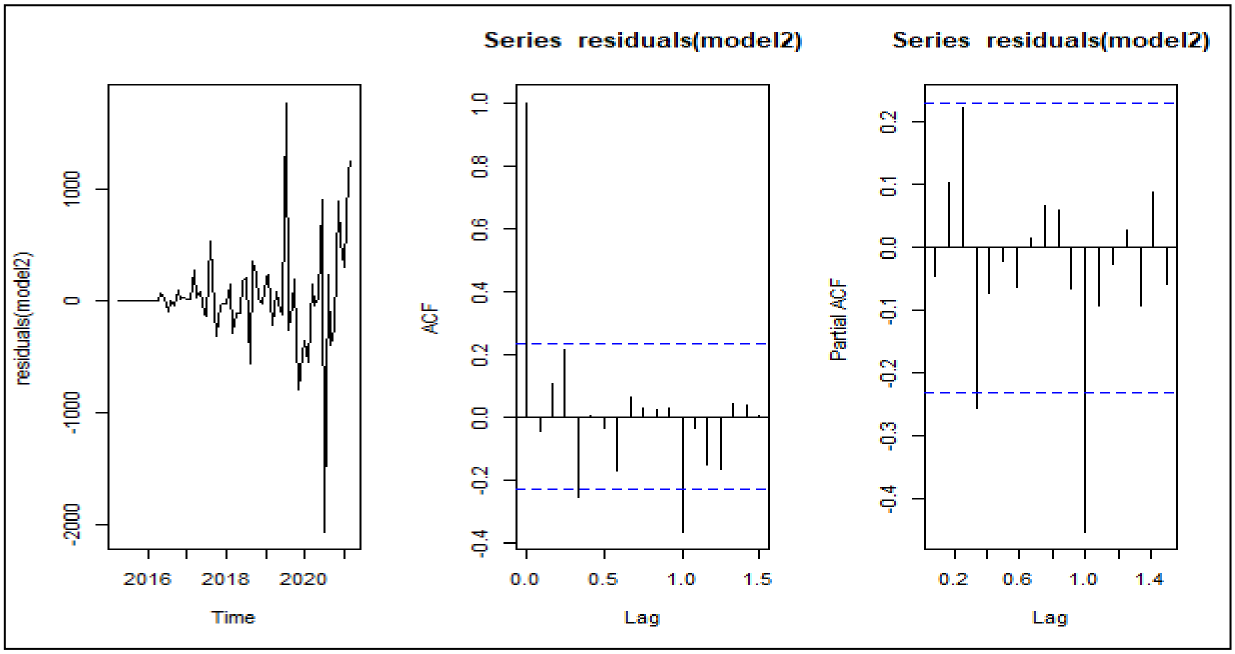
  $\varphi(B) = 1 - \varphi_1 B - \cdot\cdot\cdot -$

  $\varphi_P B^p (1 - B)Xt = Xt -$

  $Xt{-}1$

Coefficients	ar1	ma1
	-0.4894	-0.2437
s.e.	0.1536	-0.1619

sigma square estimated as 55.97: log likelihood = -202.68, AIC =411.36



- Box-Ljung test

X-squared = 0.1738, df = 1, p-value = 0.6768

SRIMA(1, 1, 4) × (0, 1, 0)12

coefficients	ar1	ma 1	ma 2	ma 3	ma 4
	-0.7608	0.0163	-0.2513	-0.0163	-0.7487
s.e	0.0930	0.1216	0.1314	0.1120	0.1222

sigma square estimated as: 4006:log likelihood = -195.7, AIC =403.39

- Box-Ljung test

X-squared = 0.15902 df = 1, p-value = 0.6901

From Ljung-Box test residuals are uncorrelated. This shows residuals are white noise and our model is good.

2.2.8 Model Evaluation and Selection :

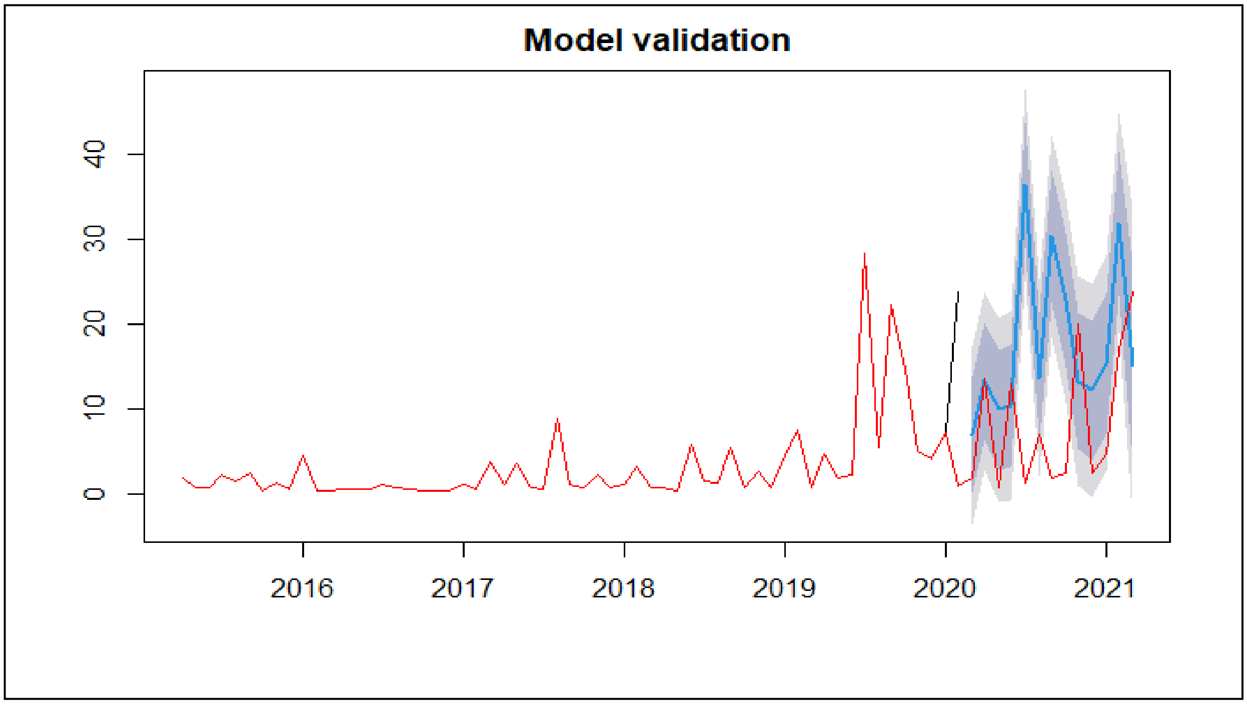
AIC of SRIMA(1, 1, 1) × (0, 1, 0) 12 is 411.36

AIC of SRIMA(1, 1, 4) × (0, 1, 0) 12 is 403.39

AIC's of both the model are almost same for both the models. But we tend to choose more parsimonious model. Hence, we go with model 1 for forecasting.

2.2.9 Model Validation:

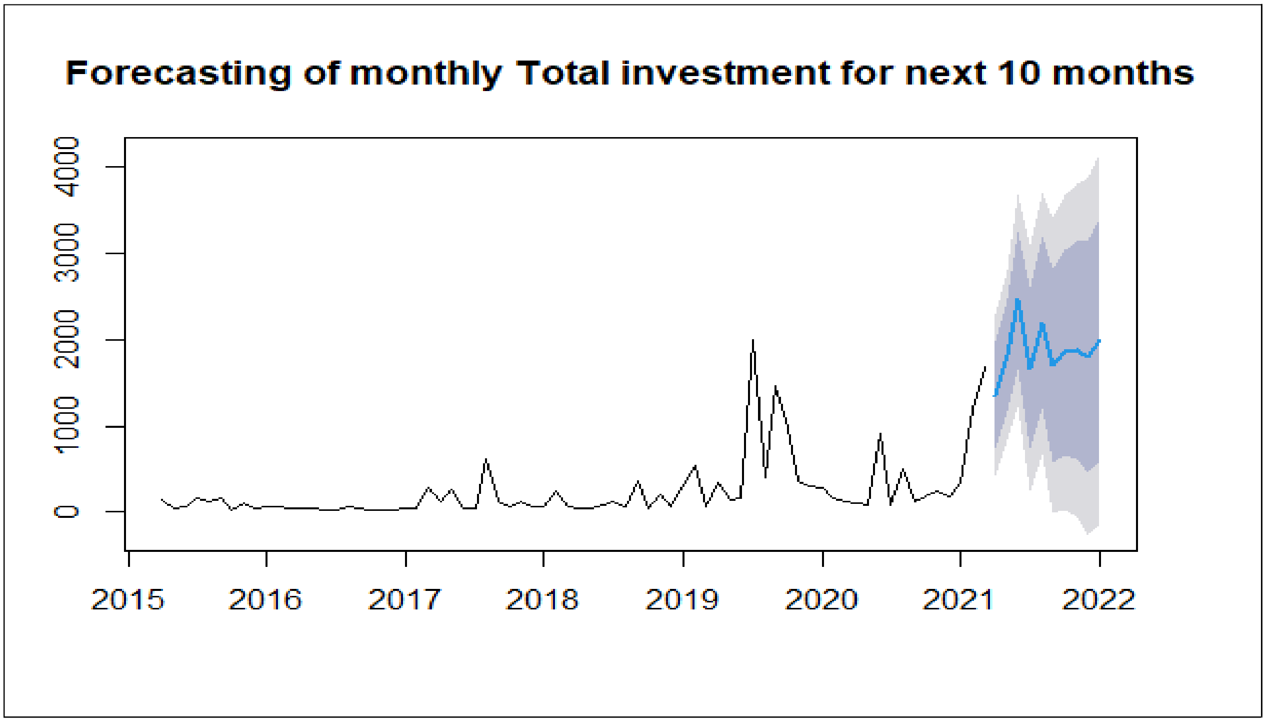
To check how good our model is, we forecast some part of data and plot them together and see whether they are considered or not.



Above plot shows our model predicts peaks and throughs very well.

2.2.10 Forecasting:

Predicting for next 10 months



As we can see, in the first quarter of 2021, there is a surge in startups funds. And also predicting our model that this trend will continue for the next 10 months. This all may happened because of the increased online transactions and more and more people using the online education system during the Pandemic era. That’s why startups like BYJU’s , and Sweggy has attracted more and more investors and will continuing the same in next upcoming months.

CONCLUSIONS:

a) The amount per investment has a step increase over the years, indicating that the investors are interested in supporting more for startups, which promises to show better performances or has a good record in the past.

b) Metro cities are most favored because of facilities and a better startup ecosystem for startups in India. Bangalore seems to have the best ecosystem for startup in India

c) The amount of fund is depending on the city, who are the investors, which type of startup.

d) Consumer internet, E-commerce, Transportation, Technology and finance are among thefive most preferred industries in terms of investment.

With that, we have done some fundamental analysis and tried to answer a few questions. We have also found some interesting patterns and trends in the Indian Startup industry in terms of city, industry vertical, etc. Further analysis can be made by combining this data with external knowledge about the startup ecosystem in India, which can lead to even better insights and trends.

REFERENCES

[1] The Indian Startup Ecosystem: Drivers, Challenges and pillars of

[2] Support by SABRINAKORRECK

[3] Data Analysis of Startups investments and funding trends in India by Piyush Anand made by Singhal

[4] S and V Dutta (2016). As Valuations Plunge, Startups Asked to Shell Out More in Tax. available at: https://economictimes.indiatimes.com/small-biz/startups/as-valuations-plunge-startups-asked to-shell-out-more-in-tax/article show/56174345.cms

[5] Startup India Action Plan Status Report. India: Ministry of Commerce. Available at: https://www.startupindia.gov.in/content/dam/investindia/Templates/public/Status_report_on_Startup_India.pdf [Accessed 9 December 2018]

[6] Growth Pattern and Trends in Startup Funding in India. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8 (12): 3721-24.

[7] Appendix

- **Python:**

Import libraries and dataset

```
# Import Libraries
import pandas as pd
import NumPy as np
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')

# Importing dataset
dataset = pd.read_csv('startup_funding.csv')
```

Understanding the dataset

```
dataset.head()
dataset.shape
# Features data-type
dataset.info()
# Checking for Null values
(dataset.isnull().sum() / dataset.shape[0] * 100).sort_values(ascending = False).round(2).astype(str)+
‘%’
```

Data Preprocessing

```
# Fixing the faulty values in 'Date' column
dataset['Year Month'] = (pd.to_datetime(dataset['Date dd/mm/yyyy']).dt.year*100) +
(pd.to_datetime(dataset['Date dd/mm/yyyy']).dt.month)
```

Exploratory Data Analysis

1. Top 10 Startups on the basis of funds acquired

```
top_10_startups=pd.DataFrame(data.groupby('StartupName')['AmountInUSD'].sum().sort_values(asc
ending=False).reset_index().head(10))
top_10_startups['AmountInUSD']=top_10_startups['AmountInUSD'].apply(lambda x: math.ceil(x))
top_10_startups
plt.figure(figsize=(20,15))
plt.title('TOP 10 STARTUPS',fontsize=25)
plt.xticks(fontsize=19,rotation='vertical')
plt.ylabel('Amount In USD',fontsize=20)
sns.barplot(x='StartupName',y='AmountInUSD',data=top_10_startups)
plt.savefig('top10startups.png',dpi=300)
plt.show()
```

- **R-Programming**

Importing library

```
library(mice)
library(shiny)
library(DT)
library(ggplot2)
library(dplyr)
library(zoo)
library(tidyr)
library(ggrepel)
library(forcats)
library(readxl)
library(tseries)
```

```
# imputation for Subvertical on the basis of industry vertical
input=mice(D[,c(3,4)],m=1,method='rf',maxit=1)
Data=complete(input,1)
D$SubVertical=Data$SubVertical


# imputation for funding amount on the basis of industry vertical,
# date, city location , funding type , etc.


input1=mice(D[,c(1,3,5,7,8)],m=1,method='rf',maxit=1)
Data1=complete(input1,1) D$AmountInUSD=Data1$AmountInUSD

sapply(D,FUN=function(x)sum(is.na(x)))summary(D)
```

Graphical

```
1) library(dplyr)
d=D %>% select(CityLocation) %>% group_by(CityLocation) %>% summarise(count=n())
%>% mutate(perc=round((count/sum(count))*100)) %>%arrange(desc(count))
d$perc=paste(d$perc,"%") # add % to the column
```

```
ggplot(head(d,10),aes(reorder(CityLocation,count),count,fill=CityLocation))+geo
m_bar(stat="identity")+theme(legend.position="none")+labs(x="",y="no. of
Startups",title="Preferred Investment Locations")+geom_label(aes(label=perc))+coord_flip() #
visualize along with %

2) d1=D %>% select(IndustryVertical,AmountInUSD) %>%
group_by(IndustryVertical) %>%
summarise(mean=round(mean(AmountInUSD),2)) %>%

arrange(desc(mean))


ggplot(d1[1:10,],aes(area=mean,fill=IndustryVertical))+


geom_treemap()+geom_treemap_text(aes(label=paste('Rs.',mean,'Cr','\n',IndustryV ertical)))+

  theme(legend.position=0) + theme(plot.title = element_text(size = 17,

    face = "bold"), legend.position = "none") +labs(title = "Industries With HighAverage
investment")
3) d2=D %>% select(InvestmentType) %>% group_by(InvestmentType) %>%
summarise(count=n()) %>%

    mutate(perc=round((count/sum(count))*100,2)) %>% arrange(desc(count))

d2$ymax=cumsum(d2$perc) #for donut

d2$ymin=c(0,head(d2$ymax,n=-1))

d2$InvestmentType=factor(d2$InvestmentType,levels=d2$InvestmentType)

ggplot(head(d2,10),aes(fill=InvestmentType,ymax=ymax,ymin=ymin,xmax=10,x
min=5))+geom_rect(color="black")+coord_polar(theta="y")+xlim(c(0,15))+geom
_label_repel(aes(label=paste(InvestmentType,round(perc),"%"),x=15,y=(ymax+y
min)/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 =
"InvestmentType")+labs(fill="InvestmentType")+guides(fill=guide_legend(keywi
dth=1,keyheight=1)) + theme(plot.title = element_text(size = 17,

    face = "bold")) +labs(title = "Donut Chart For Investment Type")
```

Kruskal-Wallis test

```
pvalue=sapply(D,FUN=function(x)kruskal.test(AmountInUSD ~ x,data =
D)$p.value)

pvalue
```

Time Series

```
# Plot for the year-month trend

D$Monyr=as.yearmon(D$Date,format='%y-%m')

D %>% group_by(Monyr) %>% summarise(Investments=n()) %>%                    # count
bymonth,year

ggplot(aes(Monyr,Investments,group=1))+geom_line(color="red")+labs(x="Year and Month
of Investment",y="Total Investment Count",title="Investment trend over the
years")+scale_x_yearmon()


#Quaterly trend
D$Qtr=as.yearqtr(D$Date,format="%y-Q%q")

D %>% group_by(Qtr) %>% summarise(Investment=n()) %>%

ggplot(aes(Qtr,Investment,group=1))+geom_line(color="blue")+labs(x="Year and
Quarter",y="Total Investment Count",title="Investment trend over the
Quarter")+scale_x_yearqtr(format="%Y-Q%q")
```

How Funding ecosystem changes with time

```
D %>% group_by(Monyr) %>% summarise(monthlyfunding = mean(AmountInUSD)) %>%
## Averages of funding over months

  ggplot(aes(Monyr,monthlyfunding))+geom_line(color='red')+

  labs(x='Time in month',y='Amount',title = 'Time series Plot')
```

Modelling time series and forcasting

We make monthly total investment as ts object for further analysis

```
TData=D %>% group_by(Monyr) %>% summarise(monthlyfunding =
mean(AmountInUSD))

series=ts(TData$monthlyfunding,start = c(2015,4),frequency = 12)plot(decompose(series))
```

Identify Level of Differencing Required

```
plot(diff(series,1))
adf.test(diff(series,1))
***Identifying the AR/MA(p/q) and Seasonal AR/MA(P/Q) components***
par(mfrow=c(1,2))
acf(diff(series,1))
pacf(diff(series,1))

model=arima(series,order =c(1,1,1),seasonal = list(order=c(0,0,0)))model
par(mfrow=c(1,3))
plot(residuals(model))
acf(residuals(model))
pacf(residuals(model))
Box.test(residuals(model),type="Ljung-Box")


***Fitting the model***


model1=arima(series,order =c(1,1,4),seasonal = list(order=c(0,1,0)))model1
par(mfrow=c(1,3))
plot(residuals(model1))
acf(residuals(model1))
pacf(residuals(model1))
Box.test(residuals(model1),type="Ljung-Box")

model2=arima(series,order =c(1,1,1),seasonal = list(order=c(0,1,0)))model2
par(mfrow=c(1,3))
plot(residuals(model2),main="series diff(series ,1)")
acf(residuals(model2),main="series diff(series ,1)")
pacf(residuals(model2),main="series diff(series ,1)")
Box.test(residuals(model2),type="Ljung-Box",main="series diff(series ,1)")
```

Model Validation

#To check how good our model is , we forecast some part of data and plot themtogether and see wether they are coincide or not.

```
library(forecast)
fit_predicted=arima(ts(TData$monthlyfunding[-c(59:71)],start =c(2015,4),frequency = 12),
                        order =c(1,1,1),seasonal = list(order=c(0,1,0),frequency=12))
forecast_pred=forecast(fit_predicted,h=13) plot(forecast_pred,main='Model
validation')
lines(series,col='red')
```

Prediction for next 10 months

```
future_pred_values=forecast(model1, h=10)
plot(future_pred_values,main='Forecasting of monthly Total investment for next10
months')
```

Dummy Variables

```
library(caret)
dummy=dummyVars(~IndustryVertical+CityLocation+InvestmentType,D,fullRan k = TRUE)
DataDu=data.frame(predict(dummy,D),AmountInUSD=D$AmountInUSD,Date=D
$Date)

DataDu$AmountInUSD=(DataDu$AmountInUSD)^lambda
library(caTool
s)
set.seed(123)
sample = sample.split(DataDu$AmountInUSD, SplitRatio =
.75)train = subset(DataDu, sample == TRUE)
test  = subset(DataDu, sample == FALSE)
```

```
model=lm(AmountInUSD ~. ,
                data=train)
R2(predict(model,test[,-118]),test$AmountInUSD)
RMSE(predict(model,test[,118]),test$AmountInUSD
)
```