SHIVAJI UNIVERSITY KOLHAPUR

DEPARTMENT OF STATISTICS

A Project Report Entitled,

"Analysis of Indian Start-up Funding Ecosystem"

Submitted By:

Dipak D. Patil

Karan P. Mane

Omkar P. Patil

Pravinsinh B. Patil

Project Guide:

Dr. . H.V. Kulkarni

CERTIFICATE

This is to certify that the Project report entitled **“A Study on Employment Attrition and Performance”** being submitted by **Dipak D. Patil, Karan P. Mane, Omkar P. Patil, Pravinsinh B. Patil**.

As partial fulfilment for the M.Sc. - II in Statistics of Shivaji University of Kolhapur is a record of bonafide work carried out by him under my supervision and guidance. To the best of our knowledge and belief, the matter presented in this project report is original and has not been submitted elsewhere for any other purpose.

Dr. H. V. Kulkarni

Head,
Department of Statistics

Place: Kolhapur

Date:

ACKNOWLEDGEMENT

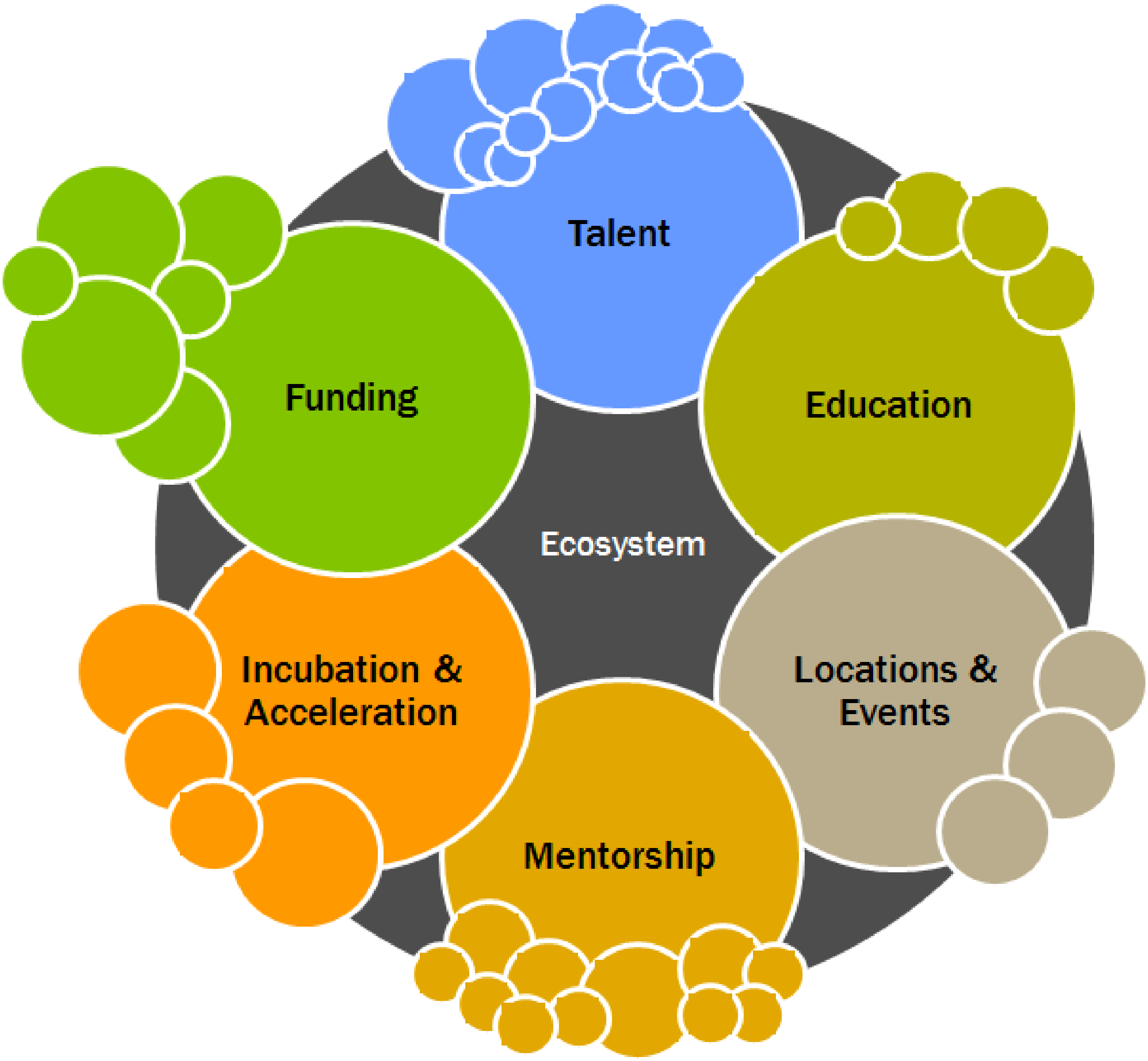
We have taken efforts in this survey however it would not have been possible without the kind of support & help of many individuals. We would like to extend our sincere thanks to all of them.

It's our great pleasure to express our sincere thanks with deep sense to Dr. H. V. Kulkarni for their valuable guidance & constant encouragement during the course of work & for providing us necessary facilities as well as for providing necessary information regarding the survey. Also, we would like to thank all non-teaching staff of our department for their help & Co-operation.

We thank all teachers for spending their valuable time for our data collection. We thank all my friends and research students for their co-operation and help which we received from them during the work throughout.

CONTENTS

Indian
START UP
Ecosystem


Talent
Funding
Education
Ecosystem
Incubation & Acceleration
Locations & Events
Mentorship

1. INTRODUCTION

A startup is a newly established company whose corporate form allows a 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 want to 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 in Mumbai and the National Capital Region (NCR), as well as some smaller cities. Further, the study investigates how the startup ecosystem has developed over the years and describes where and which kind of support is available.

1.1 Motivation

India has the 3rd largest startup ecosystem in the world .In recent years, startups have been receiving increased attention in many parts of the world. In India, the number of startups has increased fast and more support has become available in all dimensions. This project analyses the current state of the Indian startup ecosystem and has following 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?

1.2 Data Collection

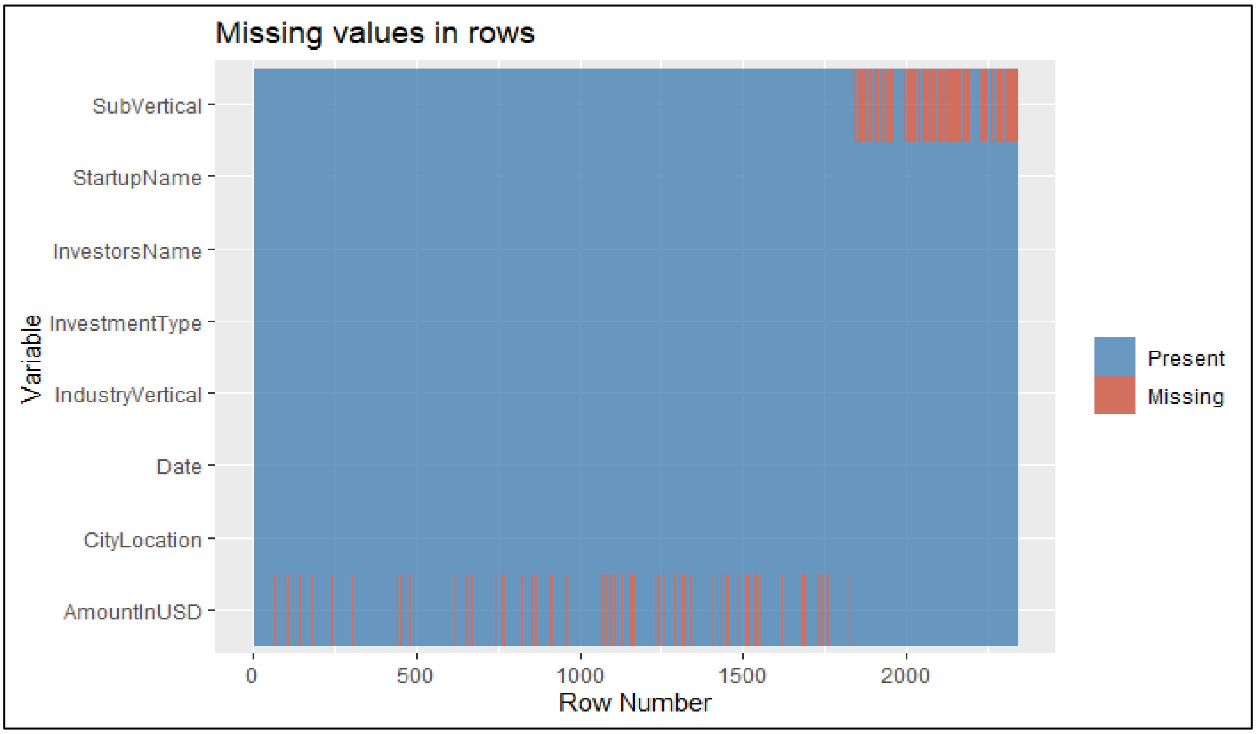
To understand and analyze the startup funding ecosystem in India, data which is secondary in nature have been collected from various websites during the period of March 2015- April 2021.

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

2. VARIABLE DESCRIPTION

- Startup Name: Name of the Company
- 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.
- City Location: City in which startup was started.
- Investors Name: Name of the investors who invested in startups.
- 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.
- Amount: Invested amount by investors in Crore.

3. CLEANING OF DATA MISSING VALUE IMPUTATION



Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted and data with missing values. This data is usually not necessary or helpful when it comes to analyzing data. When collecting data from several streams and with manual input from users, information can carry mistakes, be incorrectly inputted, or have gaps. Data cleaning helps ensure that information always matches the correct fields while making it easier for analysis. MICE is a multiple imputation method used to replace missing data values in a data set under certain assumptions about the data missingness mechanism (e.g., the data are missing at random, the data are missing completely at random).

MICE is one of the commonly used package by R users. It takes care of uncertainty in missing values and assumes that the missing data are Missing at Random (MAR) that means the probability of missing value depends only on observed value and can be predicted. In MICE there are different methods for imputation according to the type of the data. We used classification and regression trees (cart) method for our data. It creates multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations. In addition, the chained equations approach is very flexible and can handle variables of varying types (e.g., continuous or binary) as well as complexities such as bounds or survey skip patterns. We have imputed most of the missing values by using MICE package in R.

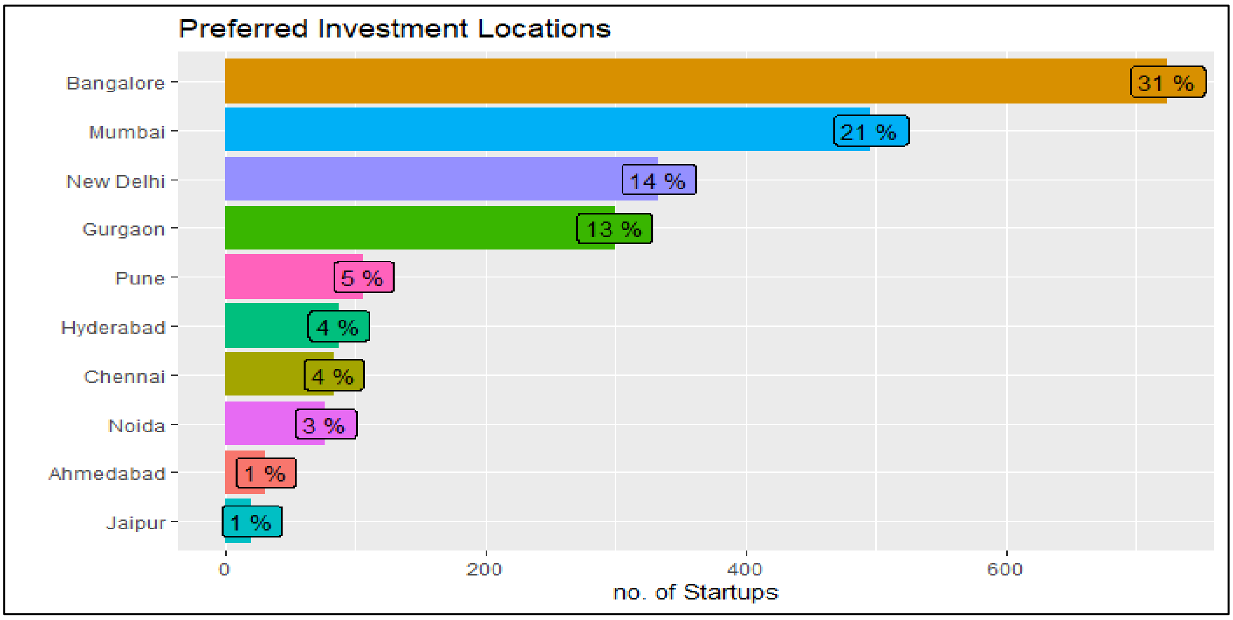
4. EXPLORATORY DATA ANALYSIS

4.1 Summary of the data

Descriptive statistics are used to describe the basic feature of the data in a study. They provide simple summaries about the sample and the measure. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

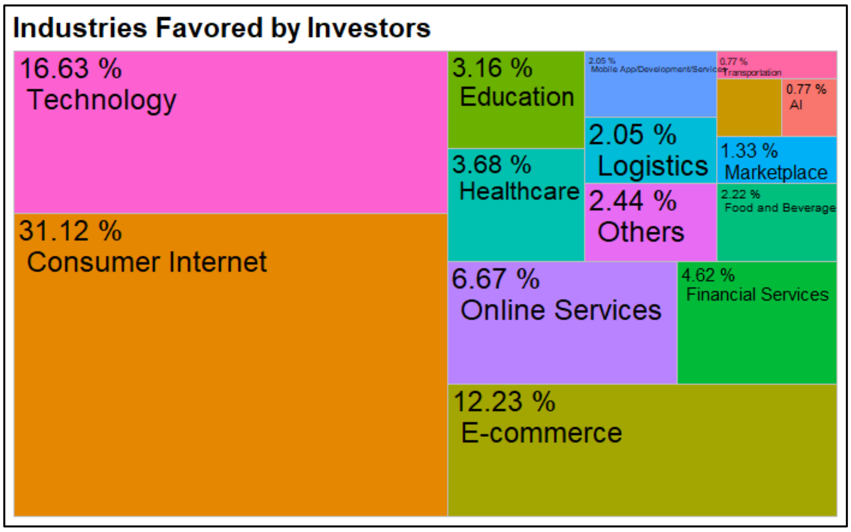
Variable Name	Amount
Min	0.0018
Median	0.20
Q1	0.050
Q3	1
Mean	2.10
Max	390

4.2 Cities Preferred by Investors



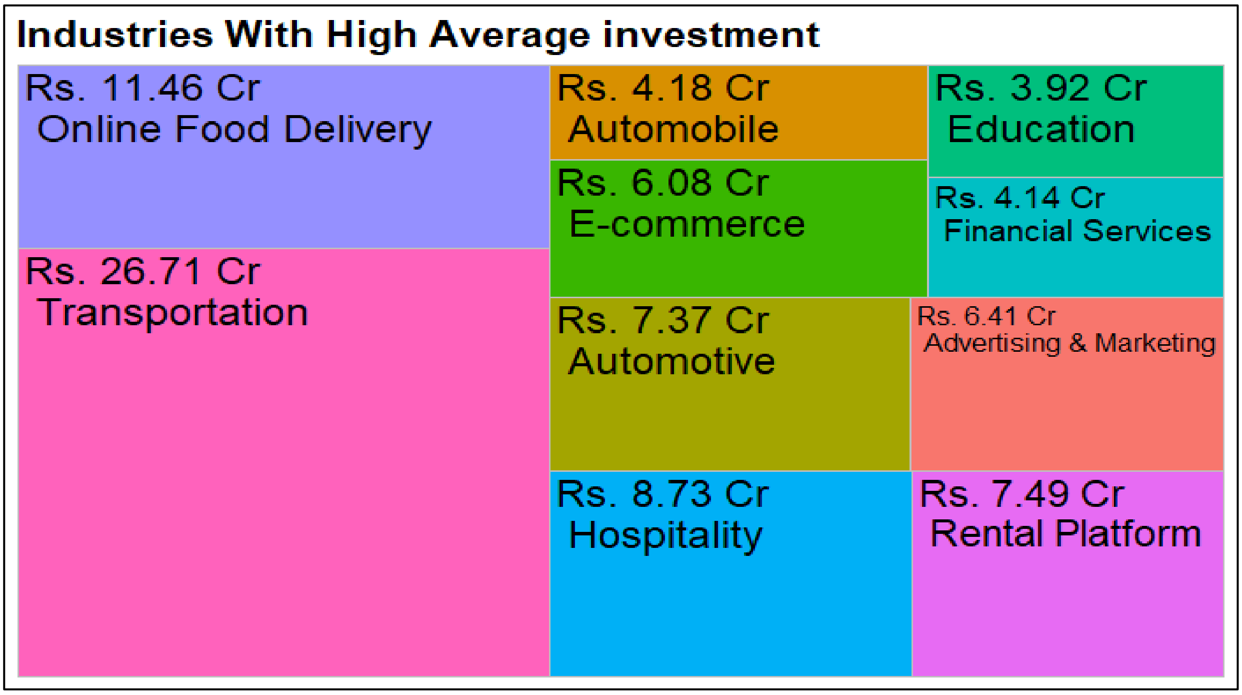
Most of the Startups are located in metropolitan cities like Bangalore, Mumbai, New Delhi, Because these are the one of the famous Global Startup hubs also Gurgaon, Pune, Hyderabad, Chennai are came under the list of emerging startups. So, the investors prefer these locations for starting their company.

4.3 Industries favoured by Investors for funding



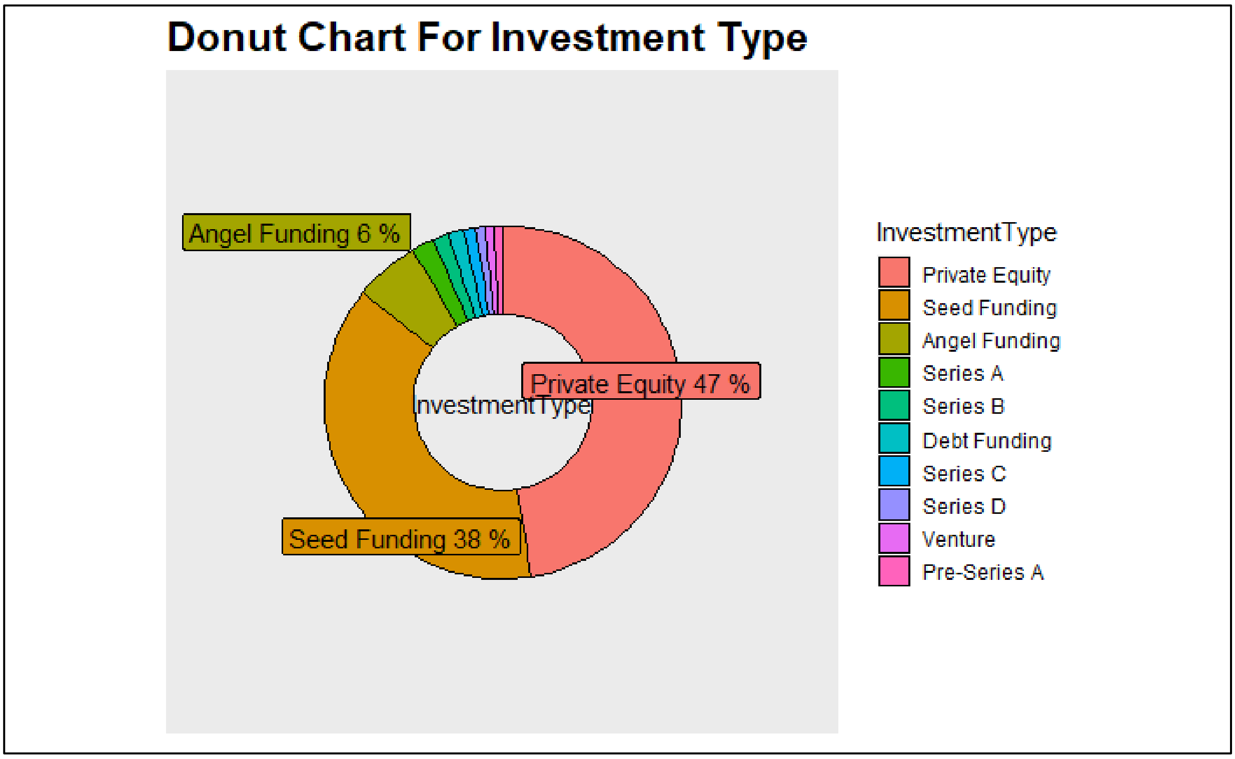
We observe that Consumer Internet related Industries are mostly favoured by Investor's. Followed by Technology, E-commerce, Online services, etc. In the above plot we see Education and Healthcare sector is also there says that these are upcoming investment sectors favoured by Investor's.

4.4 Industries with high average investment



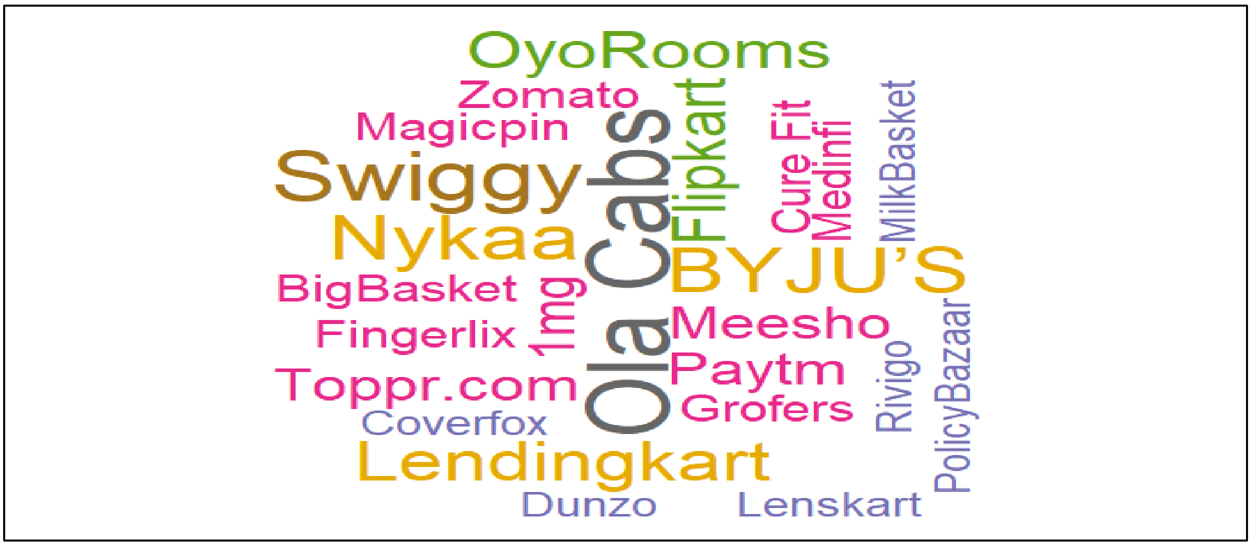
We observe from the above plot of "Indutries with High Average Investment" that Transportation, Online Food Delivery has high average investments compare to other. But, Sectors like Hospitality, Rental Platform, Automotive, Advertising and Marketing Also shows high average investment. Education, E-Commerce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

4.5 Breakdown of Investment Type



We observe that 47% of Funds are provided by Private Equity, Seed Funding and Angel Funding. Private Equity is favoured 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 wants ownership equity in the company in exchange of funds.

4.6 Which Startup has more investors confidence



We see that from wordcloud that Ola Cabs has more investor's confidence cause more no. of Investors has invested in it followed by Swiggy, BYJU's, Naykaa and so on. i.e. We see that Ola Cabs, Swiggy which are transportation and food delivery service sectors has more confidence of investors. In recent years investment in Education sector is also observes significance difference. Due to which we see BYJU's, Unacademy, Toppers.com seen in this wordcloud (We can say that Due to good network connectivity, Globalizations online learning platforms are prospering and getting good response from investors.)

4.7 Who are the important investors in Indian ecosystem

Investors Name	Total Amount	count
Softbank	431.20	6
Westbridge Capital	390	1
Microsoft, eBay, Tencent Holdings	140	1
Walmart Inc	120	1
Vijay Shekhar Sharma	100.02	2
Innoven Capital	72.17	4
Steadview Capital and existing investors	70	1
Alibaba Group, Ant Financial	68	1

We see that 'Softbank', 'Westbridge Capitals', 'Microsoft, eBay, Tencent Holdings' like investment firms as the important investors. We see Individuals name like Vijay Sharma. Here Vijay Sharma invests individually and also in collaboration with investment firms.

5. STATISTICAL ANALYSIS

5.1 Feature Selection

For getting an a more interpretable model it is better to have a selected important variables in the model. Thats 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 below0 Since P-value (=0.14) for

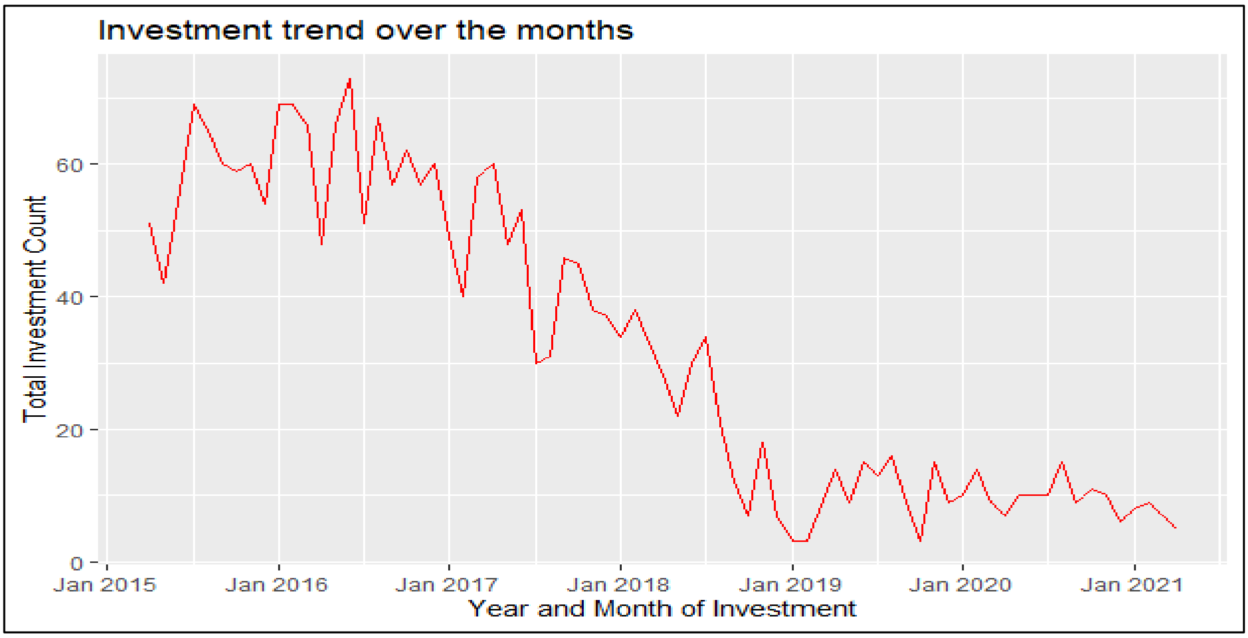
Industry vertical	Sub vertical	City Location	Investors name	Investment type
3.91*10-16	1.42*10-1	1.39*10-7	5.36*10-6	0

Since P-value (=0.14) for

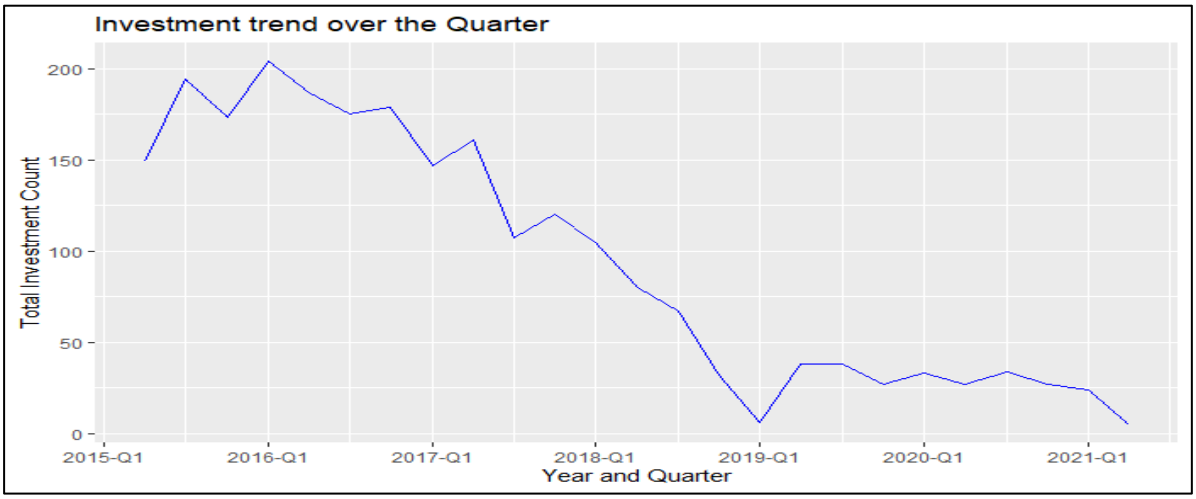
subvertical is greater than the level of significance 0.05 , therefore subvertical is insignificant with respect to the amount, we can remove it. All other are significant.

5.2 Time Series Analysis

5.2.1 Trend for number of investment over the year



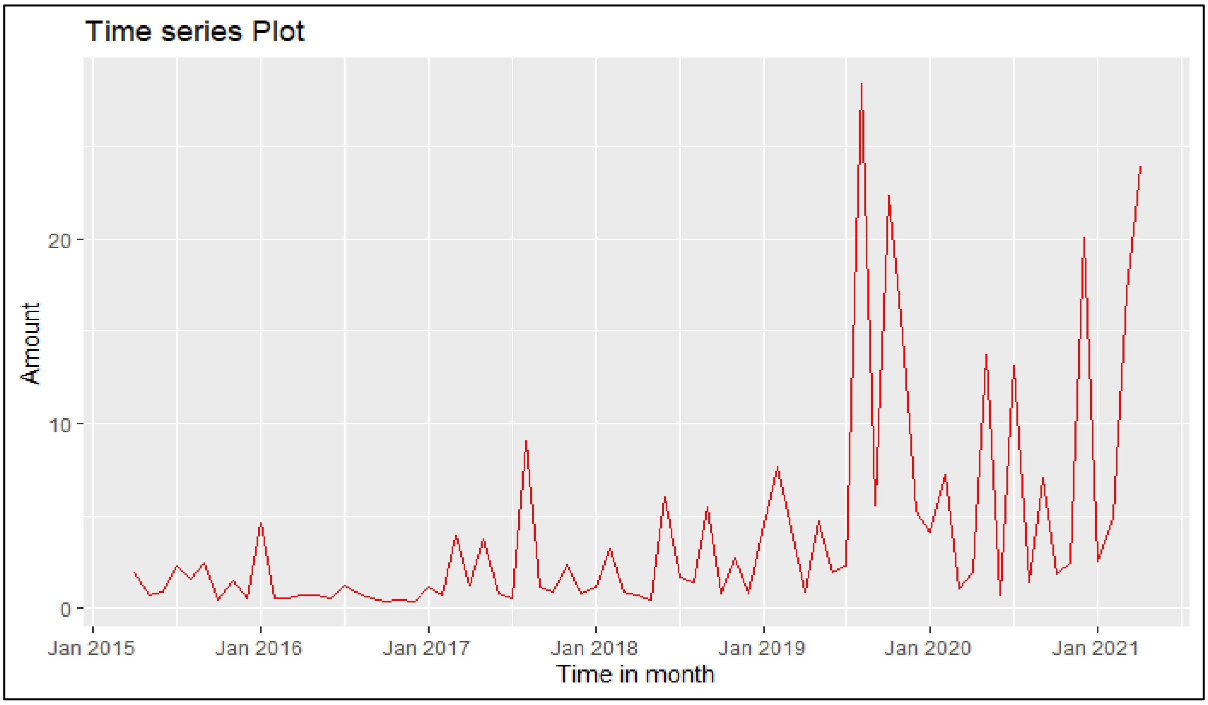
we can clearly see here a decreasing trend in the series but amount of investment is increasing.



12

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

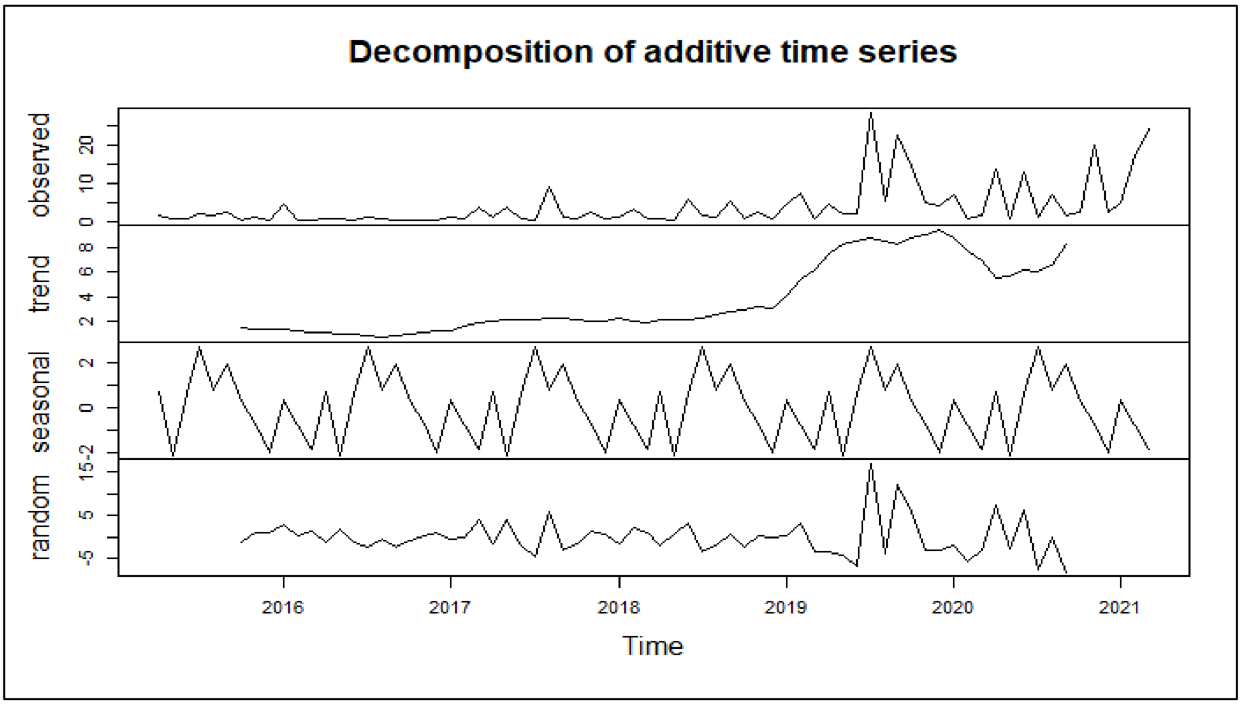
5.2.2 Monthly total investment trend



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 anlysis) 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 seed or angel funding.

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

5.2.3 Decomposition of Series



5.2.4 Tests for stationarity

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

Null hypothesis for adf test

H0 : Data is non-stationary

alternative hypothesis: stationary

Null hypothesis for PP test

H0 : No unit root (i.e. all roots are not unity)

Augmented Dickey-Fuller Test

data: series

Dickey-Fuller = -3.3255, Lag order = 4, p-value = 0.0746

alternative hypothesis: stationary

Phillips-Perron Unit Root Test

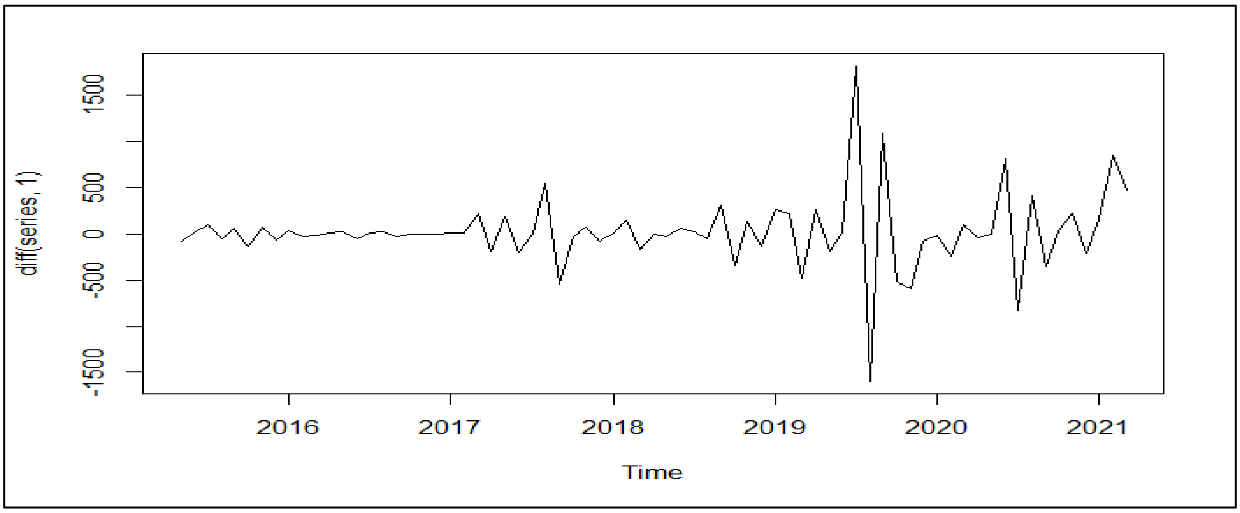
data: series

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.

5.2.5 Identification of Level of Differencing.

we tried for d=1



checking stationarity for differenced series

Augmented Dickey-Fuller Testp

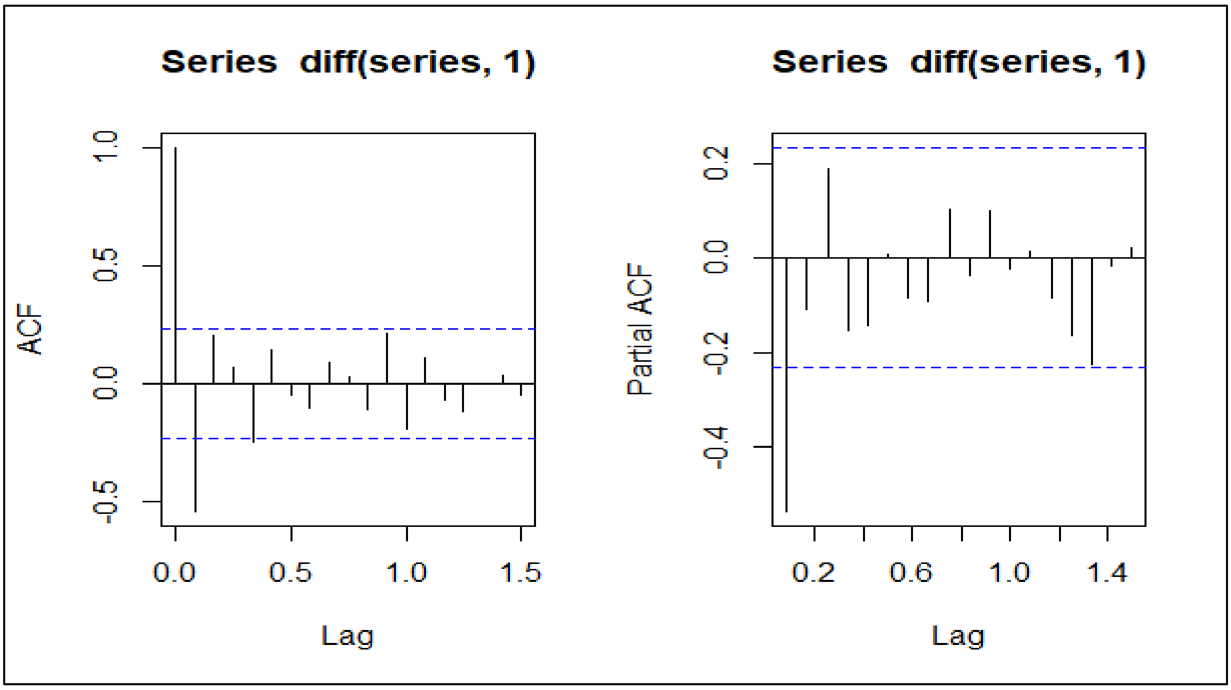
data: diff(series, 1)

Dickey-Fuller = -0.4514, Lag order = 4, p-value = 0.0124

alternative hypothesis: stationary

adf test for lag one difference shows series is stationary. so lag one difference in enough to bring stationarity.

5.2.6 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

5.2.7 Time Series Models

- Model without considering seasonality - ARIMA(1, 1, 1)

Formula:

ARIMA(p, d, q) :

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

where,

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

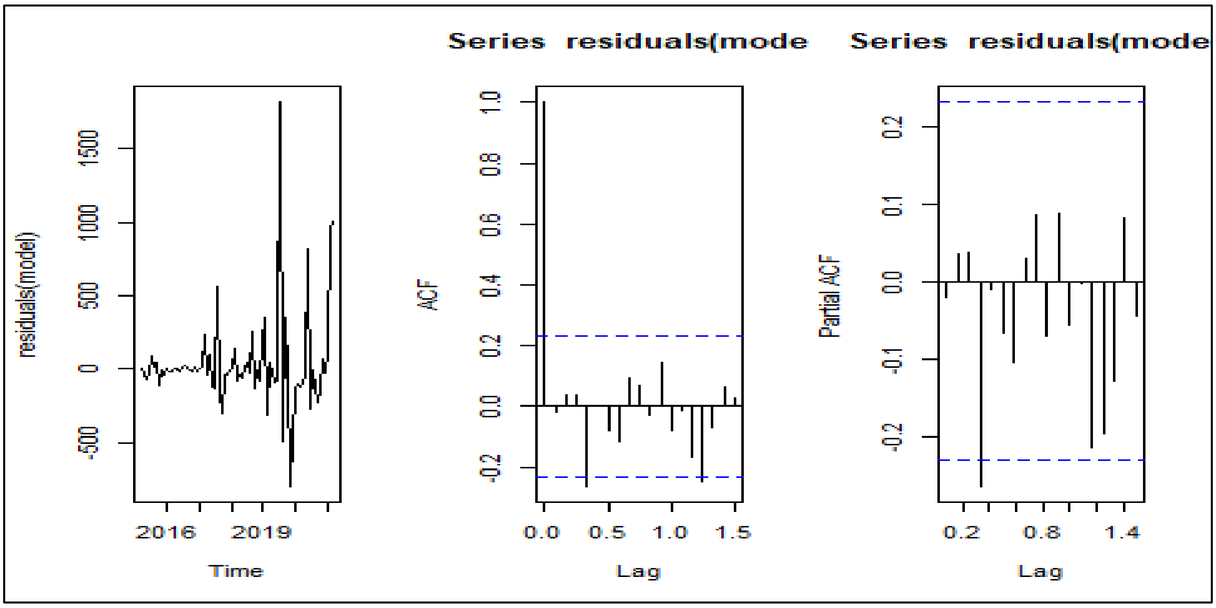
$$\varphi(B) = 1 - \varphi 1B - \cdot \cdot \cdot - \varphi pBp$$

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

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 = 0.03095, df = 1, p-value = 0.8603

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

- SRIMA(1, 1, 1) × (0, 1, 0)12

Formula:

SARIMA(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,

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

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

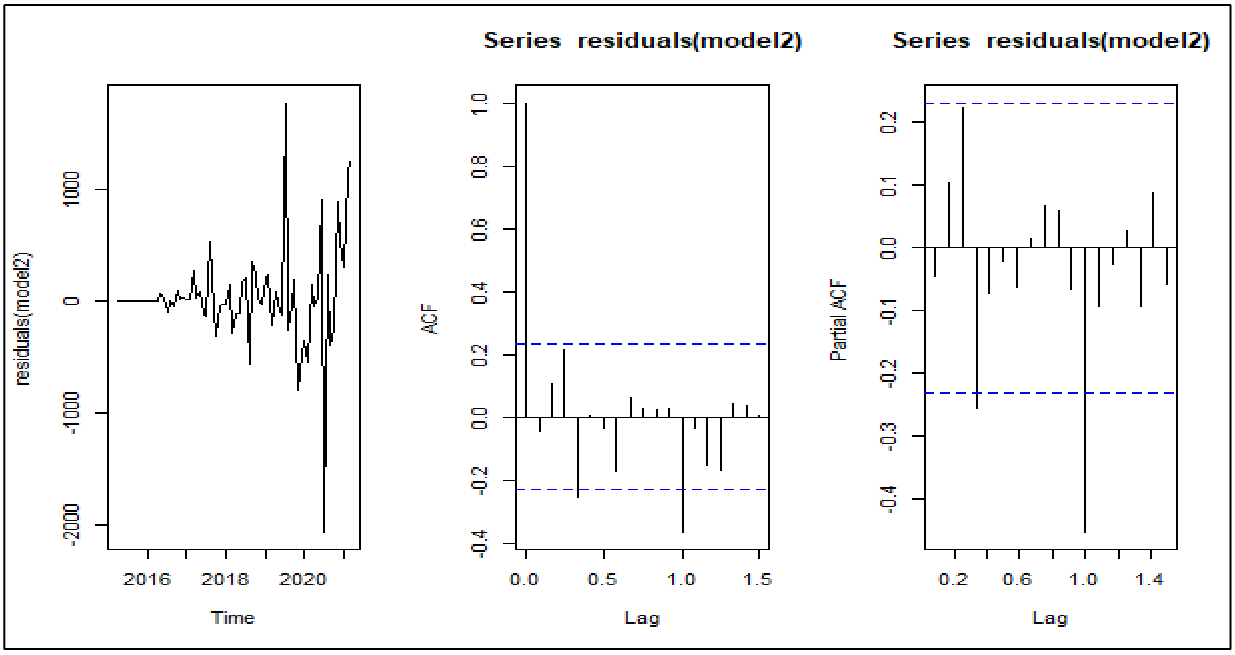
$\Phi(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdot\cdot\cdot - \Phi_P B^{P\ S}$

$\varphi(B) = 1 - \varphi_1 B - \cdot\cdot\cdot - \varphi_p B^p$

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

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



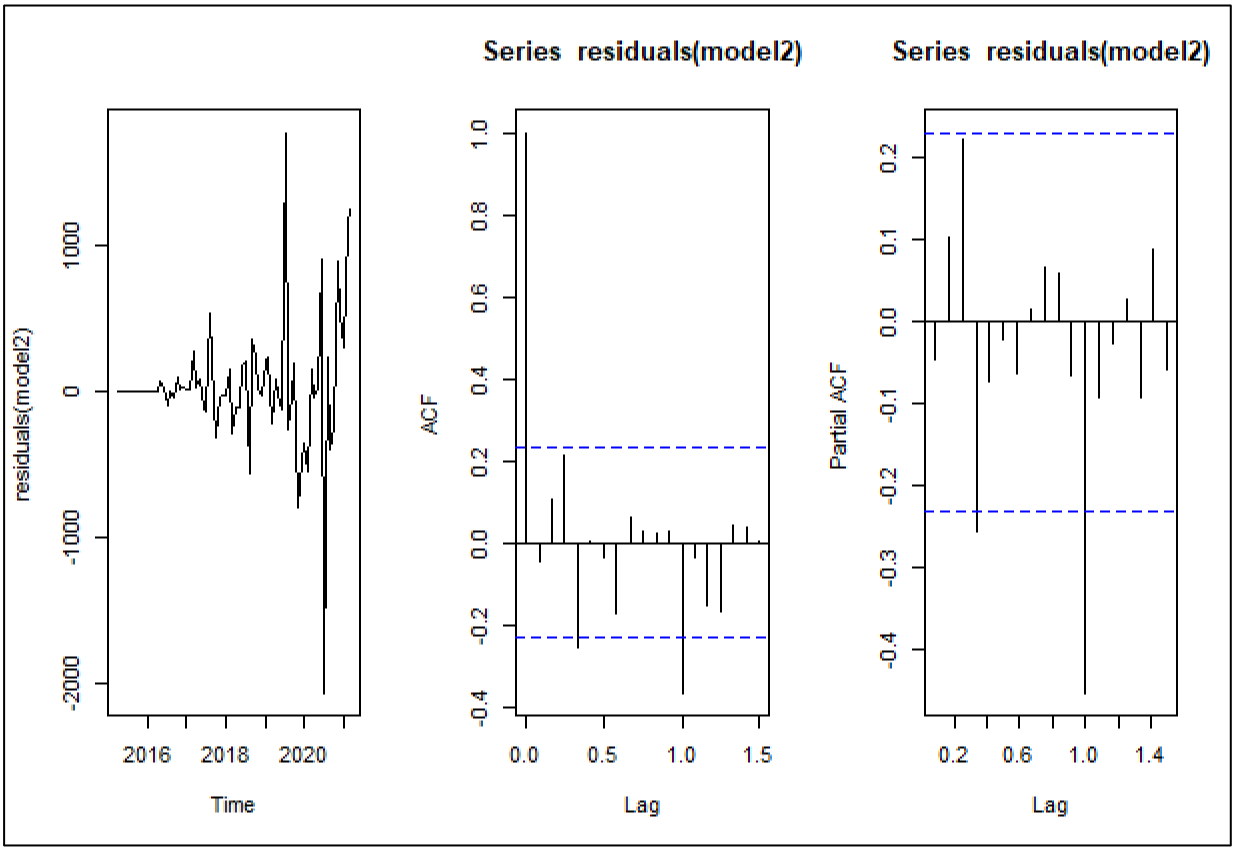
- Box-Ljung test

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

SRIMA(1, 1, 4) × (0, 1, 0)12

coefficients	ar 1	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.

5.2.8 Model evaluation and selection

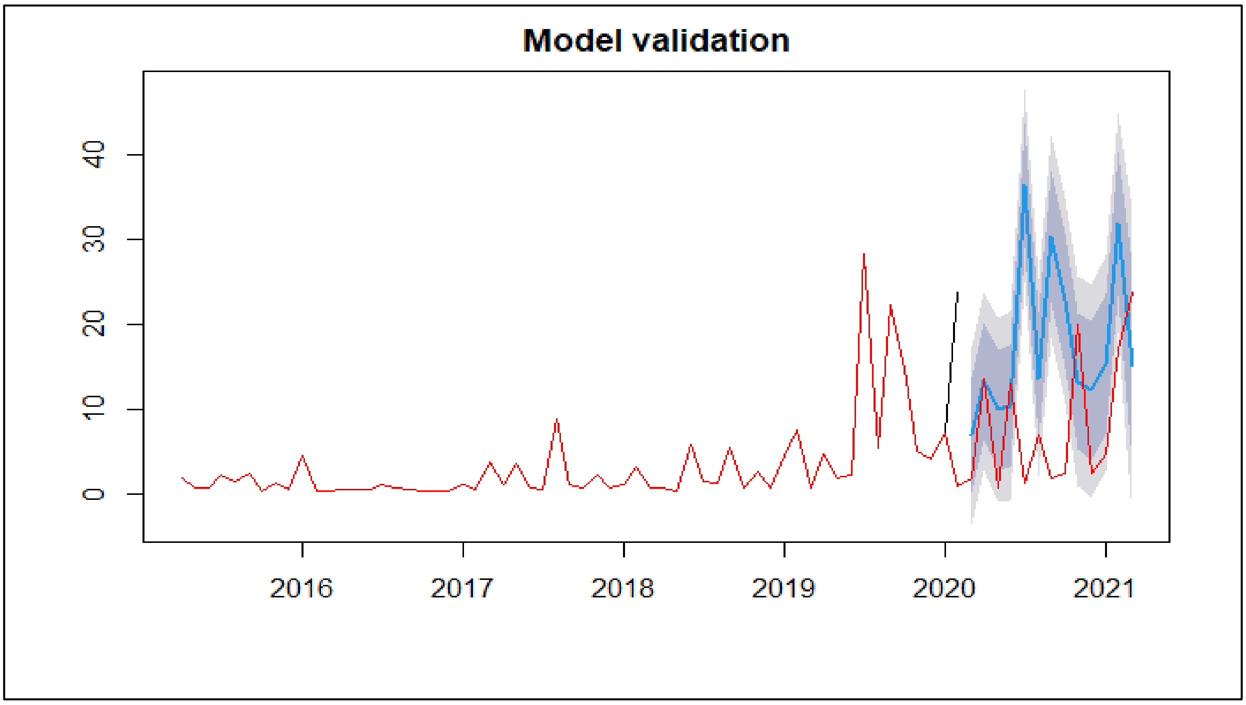
AIC of SRIMA(1, 1, 1) × (0, 1, 0) 12 is 411.36

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.

5.2.9 Model Validation

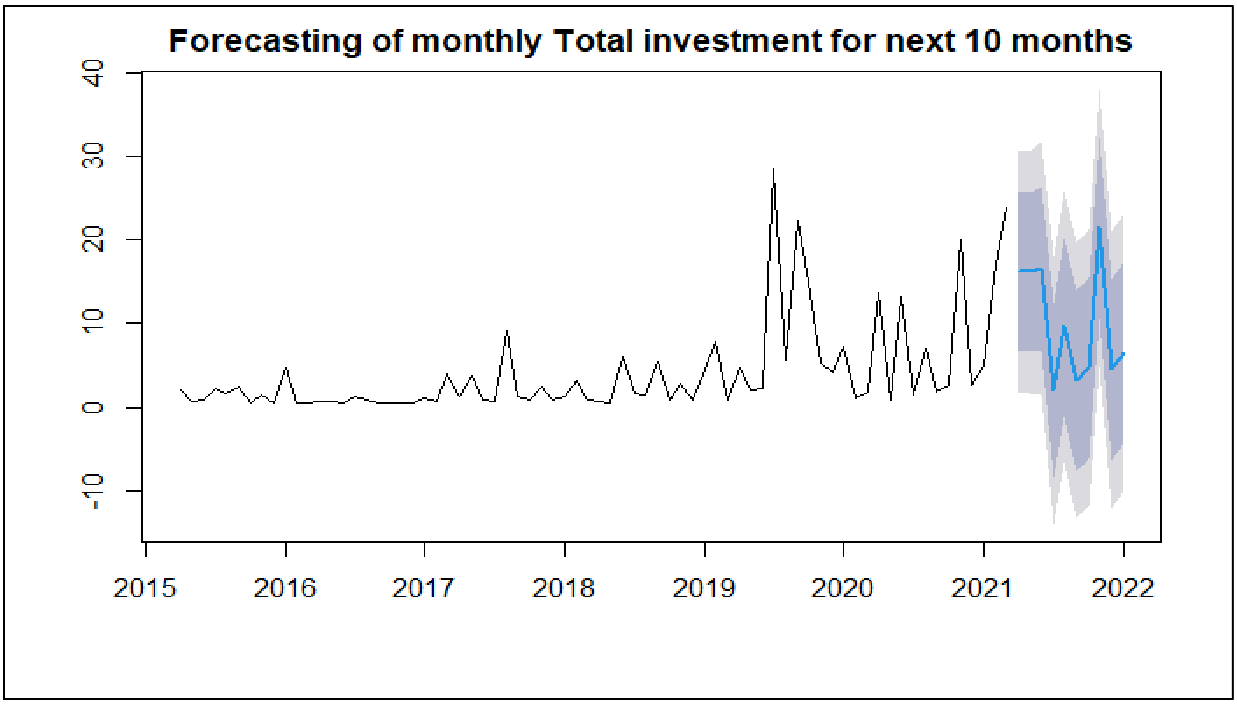
To check how good our model is, we forecast some part of data and plot them together and see weather they are coincide or not.



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

5.2.10 Forecasting

Prediction 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 Pendamic 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.

5.3 Regression Analysis

Regression analysis is used for estimating the relationship between a dependent variable and set independent variables.we used a linear regression analysis.There are four assumption associated with the regression analysis:

- Linearity
- Constant variance
- Independence of residual
- Normality of Residuals

As we already seen , here our response variable(Funding Amount) is nonnormal. To make it normal we can go for tranformation.

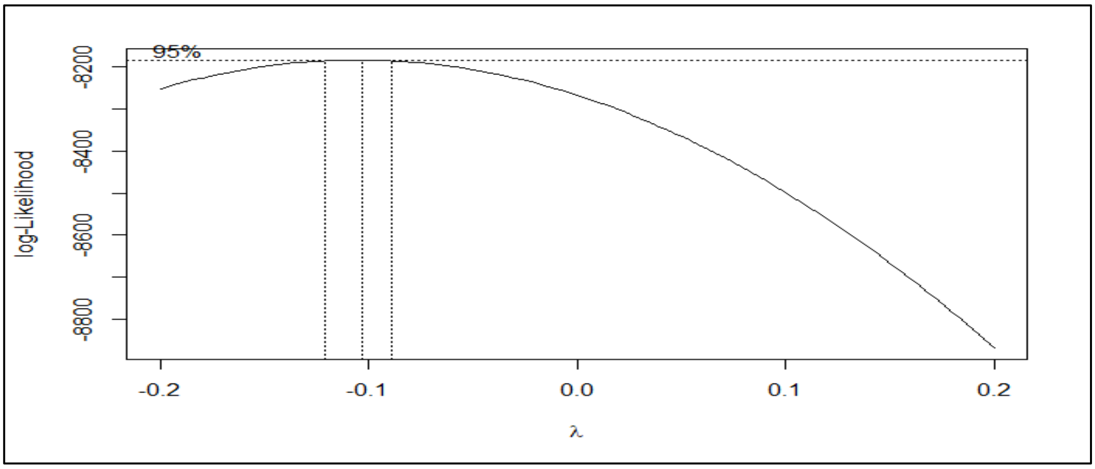
5.3.1 Power Transformation to bring normality

A Power transformation is a transformation of a non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques; if your data isn’t normal, applying a Power Transformation means that you are able to run a broader number of tests. The formula for Power transformation is If X is non-normal dependent variable and Lambda as index parameter, then transformed variable Y is given as

$$Y = \log(x) \quad ; \text{ if } \lambda = 0 ;$$

$$= x^{\lambda} \quad ; \text{ if otherwise.}$$

We found out optimize lambda =-0.13 using box-cox transformation. Then transformed amount of investment variable after transforming via power transformation looks like below,



This is not exact normal but we can proceed with assuming normality for further analysis.

5.3.2 Dummy Variables

A dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. When we want to convert Categorical data into Numeric one then dummy variables are generally used. we have categorical variable. like Industry vertical , City Location and Investment Type. All are nominal variable. So here is need to transform them. We transformed above variables into dummy variables and took these variable for model fitting.

5.3.3 Splitting the data into Train set and Test set

- Training Data set

It simply means the sample of data used to fit the model. The actual dataset that we use to train the model. The model sees and learns from this data.

- Test Dataset:

The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset. The test set is generally what is used to evaluate competing models.

Here we splitted the data in 3:1 ratio. In the training data there are 1843 observation and test set consists of 496

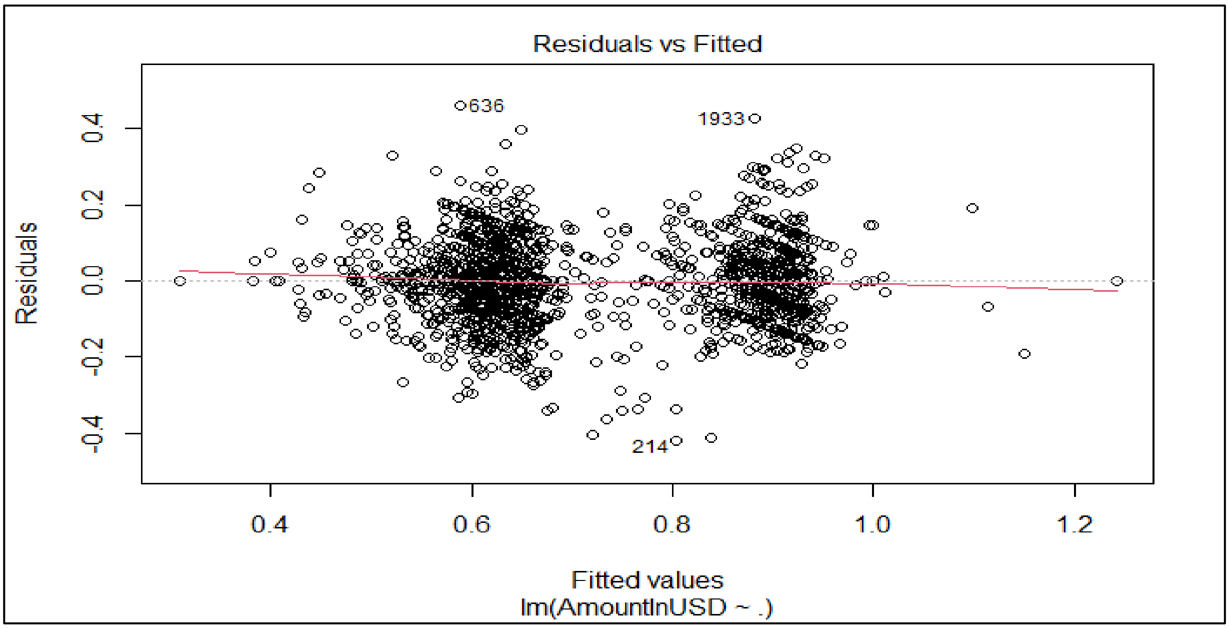
observations.

5.3.4 Regression Model

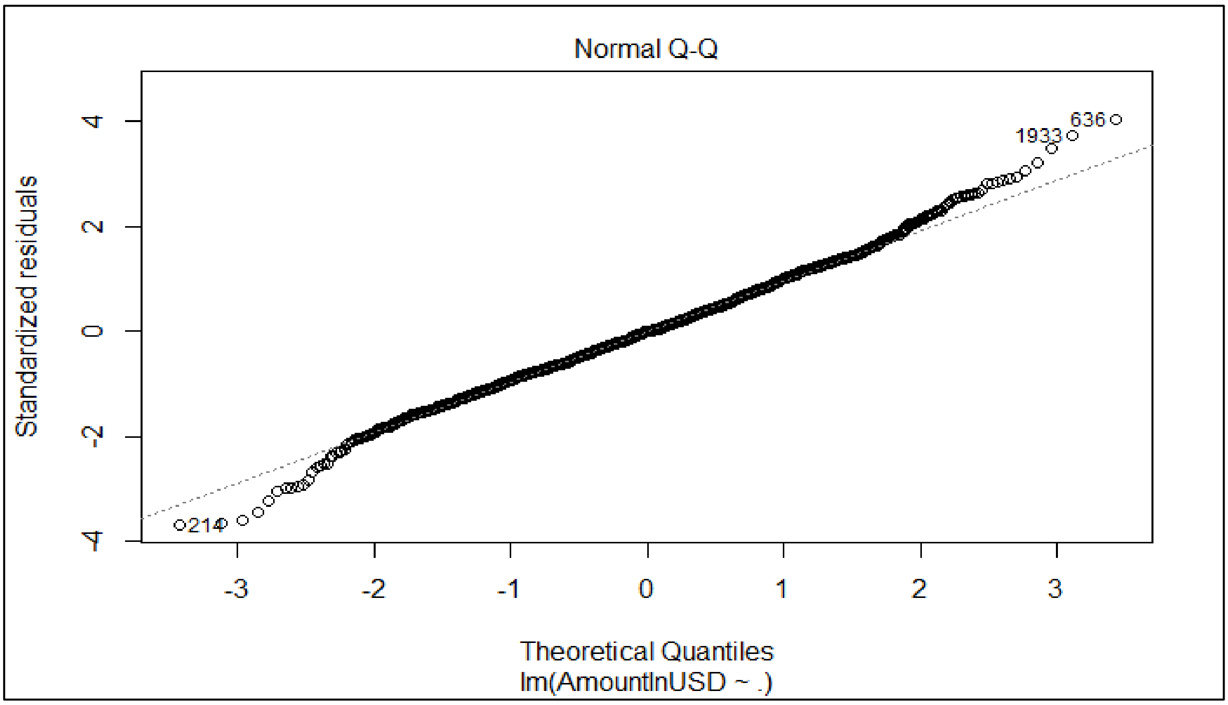
```
 model=lm(AmountInUSD . , data=train)
R2(predict(model, test[,-118]),train$AmountInUSD)
R2=0.63
```

We got coefficient of determination as 0.63. i.e. 63 % variation explained by model on training set.

5.3.5 Model Dignostics



There is no heteroscedasticity in errors i.e Variance is constant.



Errors are normally distributed.

- **<u>% of variation</u>**

Proportation of variation Explained

In statistics, the coefficient of determination, denoted Rsquare or rsquare

and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable(explained by) from the independent variable(s). Rsquare is the square of the coefficient of multiple correlation. In both such cases, The coefficient of determination normally ranges from 0 to 1.Rsquare*100 gives percentage of variation explained by independent variables in the dependent variable.

5.3.6 Leverage points

Data points which exercise considerable influence on the fitted model are called leverage points. A leverage point is a point whose x -value is distant from the other x-values. A point is a bad leverage point if its y-value does not follow the pattern set by the other data points. In other words, a bad leverage point is a leverage point which is also an outlier. A good leverage point is a leverage point which is NOT an outlier.

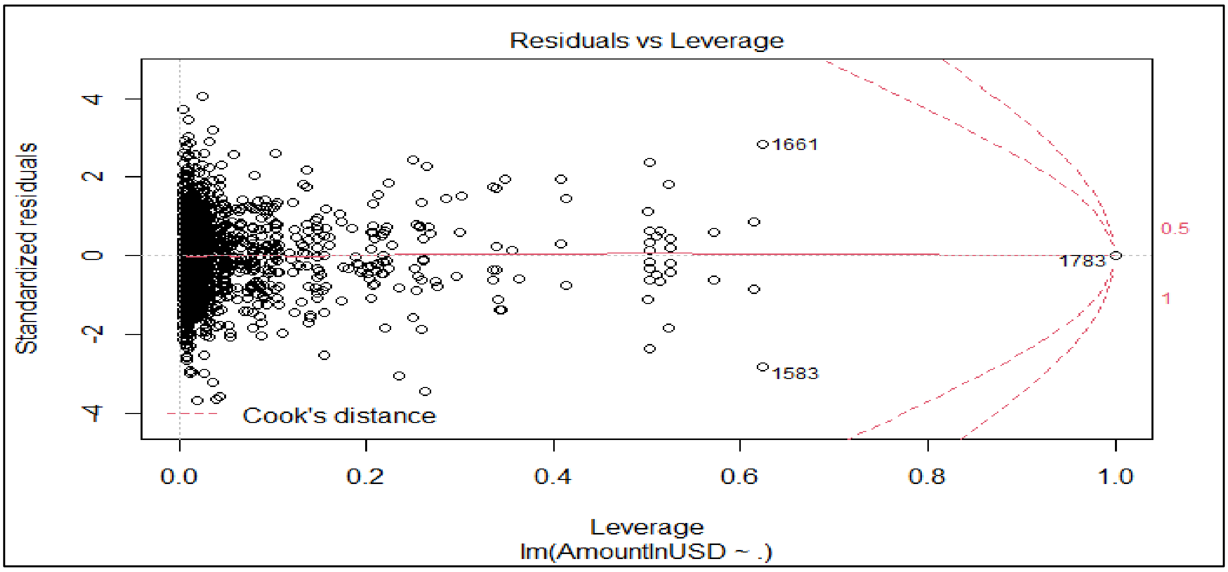
The influential point can can be identified using three methods

- Cook's D

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)' (\boldsymbol{X}'\boldsymbol{X}) \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)}{p\text{MSres}}$$

where beta(i) is the least square estimate of beta obtained after removing the ith set of observations. The ith observation is said to be influential if D(i) is greater than F(0.5,n,n-p)

i.e. Generally Di ¿ F(alpha , n, n-p) are treated as leverage points



But we haven't found leverage points in our data. So our model is now ready for the prediction of funding amount that startups would get?

5.3.7 Model Prediction

model=lm(AmountInUSD . , data=train)

R^2(predict(model,test[,-118]),test$AmountInUSD)

R^2=0.56

By Simple Linear Regression model, we got 0.56 % of accuracy in prediction

and Root mean square error by this model is 0.1158

5.4 Other Machine Learning models

5.4.1 SVM

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin).

The support Vector Regression (SVR) uses the same principles as the SVM for classification with only a few minor differences. First of all because output is a real number it becomes very difficult to predict the information at hand which has infinite possibilities. In the case of regression a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact there is also a more complicated reason the algorithm is more complicated therefore to be taken in consideration. However he main idea is always the same: to minimize error individualizing the hyperplane which maximizes the margin keeping in mind that part of the error is tolerated.

By Support Vector Machines (svm) algorithm, we got 41.51% % of variance explained in prediction and Root mean square error by this model is 0.1351.

5.4.2 Decision Tree

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. By Recursive Partitioning (rpart) algorithm, we got 55.97% % of variance explained in prediction and Root mean square error by this model is 0.1166

5.4.3 Xgboost

Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive modelling problems. Ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. This is a type of ensemble machine learning model referred to as boosting. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm.

By Extreme Gradient boosting(xgboost) algorithm, we got 58.07% % of variance explained in prediction and Root mean square error by this model is 0.1135.

5.4.4 Random forest

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression. Random forest is a bagging technique and not a boosting technique. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees.

A random forest combines the result of multiple predictions which aggregates many decision trees, with some helpful modifications.By random forest algorithm, we got 59.81% % of variance explained in prediction and Root mean square error by this model is 0.1101.

Algorithm percentage of variance explained RMSE

Algorithm	R^2	RMSE
Regression Analysis	56.78	0.1158
Random Forest	59.81	0.1101
SVM	41.51	0.1351
Recursive Partitioning	55.97	0.1166
Extreme Gradient Boosting	58.07	0.1135

In above all the Machine learning algorithm, We have nearly same Root mean square error (RMSE) which tells how concentrated the data is around the line of best fit i. e. for the given algorithm. As all gives nearly same RMSE the model which explains more variation in the test data is the best model. Here Random forest algorithm explains 59.81% of variation in the response variable which is highest among all the algorithms. SVM algorithm is not giving good results. compare to other algorithm. Also, Extreme Gradient boosting(xgboost) algorithm explains nearly 58% of variation in the response variable and that Simple linear regression model and recursive partitioning explains nearly 56% of variation in the response variable which are giving good results. But,

the best one is Random Forest algorithm.

6. CONCLUSIONS

- 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.

- 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

- The amount of fund is depends on the city, who are the investors, which type of startup.

- Consumer internet, E-commerce, Transportation, Technology and finance are among the five 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.

7. 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 Vermaand, Vikas Singhal