**Dr.D.Y.Patil Institute of Technology**

Pimpri, Pune-411018

A MINI-PROJECT REPORT

ON

**“Profit Segmentation for Investment Companies for Decision Making Process”**

Submitted by

BCOA41 Gauri Gulwane

BCOA42 Anubhuti Rane

BCOA43 Sampada Gaonkar

BCOA44 Tamanna Kasliwal

**Department of Computer Engineering**

2019-2020

**Dr.D.Y.Patil Institute of Technology**

PIMPRI, PUNE-411018

**Department of Computer Engineering**

CERTIFICATE

Certified that the mini-project work entitled **“Profit Segmentation for Investment Companies for Decision Making Process”** is a bonafide workcarried out by

**Gauri Gulwane BCOA41**

**Anubhuti Rane BCOA42**

**Sampada Gaonkar BCOA43**

**Tamanna Kasliwal BCOA44**

The report has been approved as it satisfies the academic requirements in respect of mini-project work prescribed for the course.

……………...…………………………

**Subject In-charge**

**Abstract**

Investment companies have to take major decisions regarding in which company to invest in. To make this decision, data analysis techniques are required for such tasks. The main attribute that constitutes the company as valuable are its profit margins. For this, other independent various like Research and Development, Marketing, Statistics are taken into consideration. Thus, linear relation is developed between the independent and dependant attributes. Finally, by k-means clustering, we conclude the valuability of a company for the decision-making.

**Content**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Name** | **Page No** |
| **1** | **Introduction** | **5** |
| **2** | **Implementation** | **6** |
| **3** | **Result and Analysis** | **7** |
| **4** | **Conclusions & Future Enhancements** | **10** |
| **5** | **References** | **11** |

**List of Figure**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Name** | **Page No** |
| **1** | **Visualization (Training Data)** | **7** |
| **2** | **Visualization**  **(Test Data)** | **8** |
| **3** | **The plotting by Elbow Method** | **8** |
| **4** | **K-means Clustering plot** | **9** |

**I. Introduction**

The investment on the start-up companies are increasing in number. The companies that are investing need profit from the start-up companies to generate some output that they initially loaned to them.

Thus, data analysis techniques are required where they predict the group of companies that generate highest profit. The techniques that are used are as follows:-

1. Multiple Regression
2. Linear Regression
3. K-means Clustering

The dependent attribute profit depends upon a number of factors like Research and Development, Marketing, and Sales.Thus a multiple regression class is used to fit into the dataset in order to find out which independent variable is linearly proportional to the dependant variable. For this, backward elimination method is used to remove all the noise attributes.Thus, we obtain the new dataset that fits to the linear regression model. We use visualization method to plot our data points. Lastly, we use k means clustering methods to find the cluster of companies that give the highest profit.

**II. Implementation**

Import the Dataset of 50 Start-up companies

Use categorical factorization to remove string values

Spilt the dataset into training and test dataset

Plot the cluster and conclude the result

Fit Multiple Regression class to train the dataset

Use k-means clustering find the clusters

Predict the results of the test dataset with the actual outcome

Use backward elimination for the independent attributes

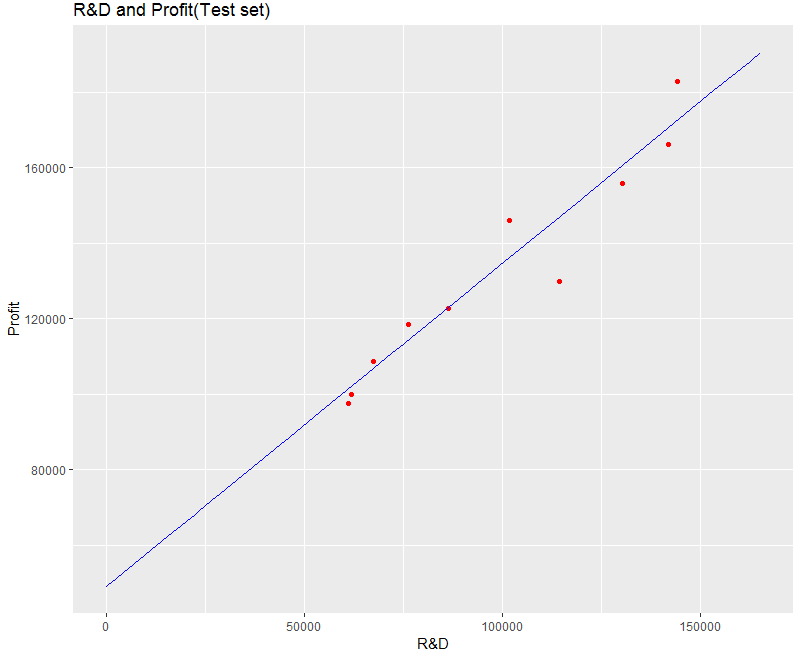
Use linear regression visualization to fit the graph

**III. Result and Analysis**

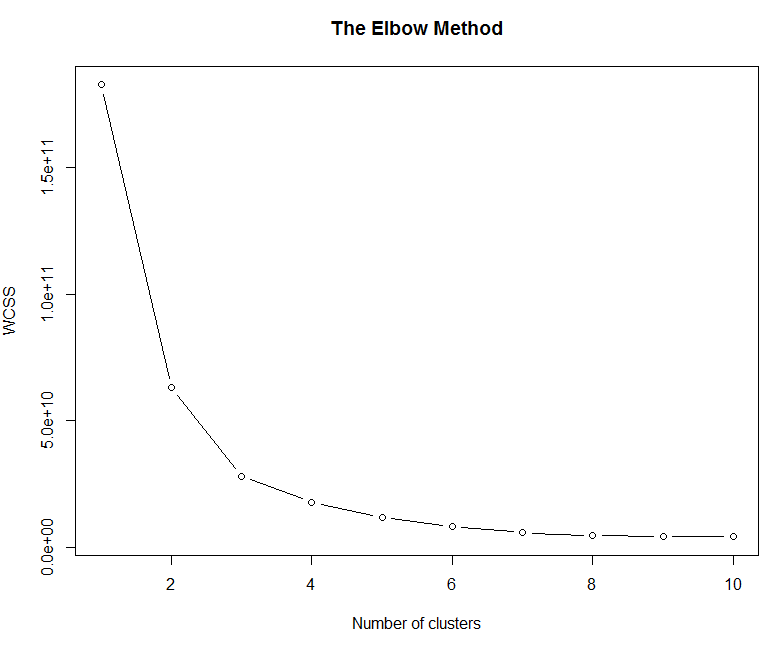
The analysis is done on the gathered data which is stored in the CSV file. For analysis purposes, we use data analysis algorithms. After analyzing the data result will also give an idea about the probability of companies that generate the highest profit.



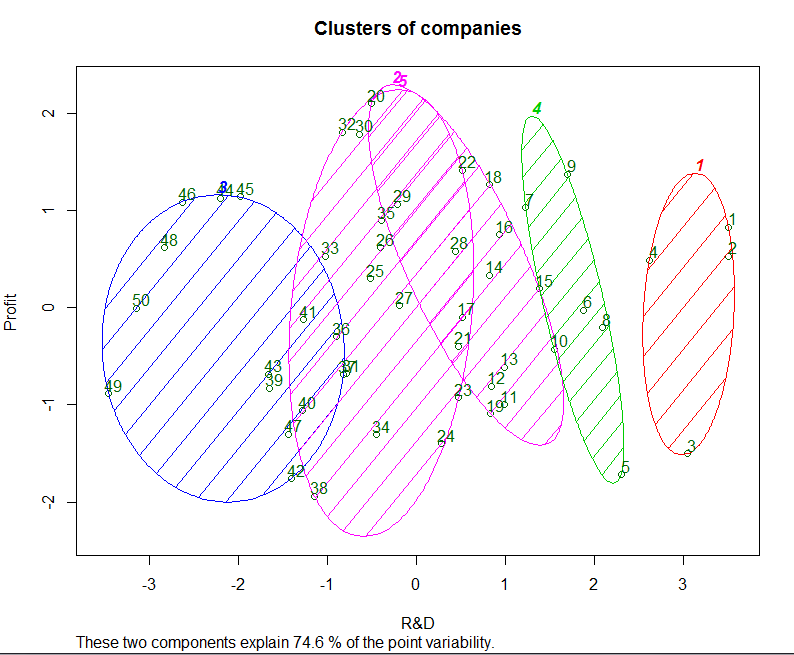
Fiq 1 :Visualization (Training Data)



Fiq 2 : Visualization (Test Data)



Fiq 3:The plotting by Elbow Method



Fiq 4: K-means Clustering plot

**IV. Conclusions**

Hence, using this application, the companies would identify the start-ups which have the scope of getting more profit. Identifying this will help in the decision making process hence increasing the profits of the investment companies. The generated graph showing the linear relationship between R&D and dependent attribute profit can be used to predict the test dataset.

**V. References**

1) Sonoda, S., Takahashi, Y., Kawagishi, K., Nishida, N., & Wakao, S. (2007). *Application of Stepwise Multiple Regression to Design Optimization of Electric Machine. IEEE Transactions on Magnetics, 43(4), 1609 – 1612.* Doi : 10.1109 / tmag. 2006 .892225 

2) Wang, P., Ge, R., Xiao, X., Zhou, M., & Zhou, F. (2017). hMuLab: A Biomedical Hybrid MUlti-LABel Classifier Based on Multiple Linear Regression. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(5), 1173–1180. doi:10.1109/tcbb.2016.2603507    
  
3) Ming Hua, Man Ki Lau, Jian Pei, & Kui Wu. (2009). *Continuous K-Means Monitoring with Low Reporting Cost in Sensor Networks. IEEE Transactions on Knowledge and Data Engineering, 21(12), 1679–1691.* doi:10.1109/tkde.2009.41

4)Links: Dataset-[www.superdatascience.com](http://www.superdatascience.com)

5)[www.stackflow.com](http://www.stackflow.com)