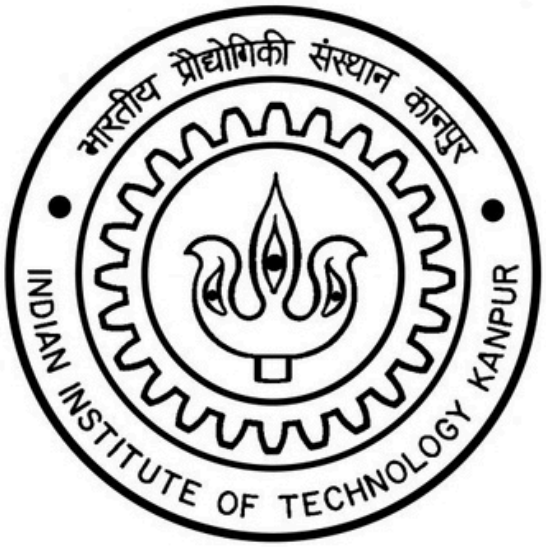
PROJECT REPORT
ON
Analytical Study of Startups

Indian Institute of Technology, kanpur



MTH 209 : Data Science lab 2

Under the guidance of :

Prof. Subhajit Dutta

Submitted By:

Rohit Kumar
Jayant Jha
Chaitanya

Acknowledgements

We extend our sincere appreciation to Prof. Subhajit Dutta for their steadfast support and invaluable guid ance throughout the implementation of this churn modeling project. Their expertise and mentorship have played a pivotal role in shaping the project's trajectory, refining analytical techniques, and deepening our understanding of customer churn dynamics. We are genuinely grateful for their commitment to academic excellence, which has inspired us to explore and analyze churn data comprehensively. Their encouragement and insights have been instrumental in the project's success, and we are truly privileged to have had the opportunity to learn under their guidance.

Contents

Chapter 1.

INTRODUCTION

Startups have become vital catalysts for innovation, job creation, and economic progress both in India and around the world. In India, the startup ecosystem has seen remarkable growth, driven by factors like favourable demographics, technological advancements, and supportive government policies. With approximately 9,000 technology-based startups, India ranks as the third-largest startup hub globally. These startups span various sectors and have contributed significantly to job creation and economic growth.

However, startups face challenges such as high failure rates and regulatory complexities. Despite these hurdles, success stories abound, with Indian startups achieving unicorn status and making their mark on the global stage.

Meanwhile, startup ecosystems worldwide are thriving, fostering innovation and disruption across industries. From Silicon Valley to Shanghai, startups have reshaped traditional business models and driven technological advancements. While the startup landscape is characterized by rapid growth and innovation, it also entails risks and uncertainties.

In this era of unprecedented technological advancement and globalization, startups in India and worldwide are poised to continue shaping economies and societies. As we delve into the world of startups, it's essential to recognize their transformative potential and enduring impact on a global scale.

1.1 STATISTICAL STUDY

Understanding the factors that contribute to the success or failure of startups is crucial for policymakers, investors, entrepreneurs, and researchers alike. Statistical studies play a pivotal role in shedding light on the dynamics of startup ecosystems, providing valuable insights that can inform decision-making and improve outcomes.

At the heart of the startup journey lies a delicate balance between risk and reward. While startups hold the promise of innovation, job creation, and economic growth, they also face

significant challenges and uncertainties. The high failure rates observed among startups underscore the importance of identifying the determinants of success and failure.

By conducting statistical analyses, researchers can uncover patterns, trends, and correlations that elucidate the underlying drivers of startup outcomes. These analyses can examine a wide range of factors, including market conditions, industry dynamics, funding sources, management practices, and regulatory environments.

Insights gleaned from statistical studies can inform various stakeholders. For entrepreneurs, understanding the factors associated with successful startups can guide strategic decision-making, resource allocation, and risk management. Investors can use statistical analyses to identify promising investment opportunities, mitigate risks, and optimize portfolio returns. Policymakers can leverage statistical findings to design policies and programs that foster a conducive environment for startup growth and innovation.

Moreover, statistical studies contribute to the body of knowledge surrounding entrepreneurship, enriching academic research and advancing theoretical frameworks. By rigorously analyzing data and testing hypotheses, researchers can deepen our understanding of the mechanisms that drive startup success and failure, paving the way for evidence-based policymaking and practical interventions.

In the context of the Indian and global startup ecosystems, conducting statistical studies takes on added significance. As these ecosystems continue to evolve and mature, the need for data-driven insights becomes increasingly pressing. By systematically analyzing startup data, researchers can identify best practices, emerging trends, and areas for improvement, ultimately contributing to the resilience and sustainability of startup ecosystems.

In summary, statistical studies on startup success and failure are essential for elucidating the complex dynamics of entrepreneurship, informing decision-making, and driving innovation and growth. By leveraging data and analytical techniques, stakeholders can unlock valuable insights that have the potential to shape the future of startup ecosystems both in India and worldwide.

1.2 RESEARCH QUESTIONS AND OBJECTIVE

1.2.1 OBJECTIVE

In the study, the objective is to investigate the factors influencing the success or failure of startups within a particular ecosystem, such as India's startup ecosystem. The research questions typically revolve around understanding the determinants of startup success, identifying patterns or trends among successful and failed startups, and exploring the impact of various factors such as funding sources, industry sector, geographic location, management practices, and regulatory environment on startup outcomes. These questions aim to provide insights that can inform decision-making by entrepreneurs, investors, policymakers, and other stakeholders involved in supporting startup growth and innovation.

1.2.2 RESEARCH QUESTIONS

1. What are the key determinants of startup success within the chosen ecosystem?
2. How do various factors, such as funding sources, industry sector, and geographic location, influence the performance and outcomes of startups?
3. What are the common patterns or trends observed among successful startups, and how do they differ from those of failed startups?
4. To what extent do management practices, regulatory environment, and external market conditions impact startup outcomes?
5. What implications do the findings of the study have for entrepreneurs, investors, policymakers, and other stakeholders involved in supporting startup growth and innovation?

These research questions aim to guide the investigation into the multifaceted aspects of startup success and failure within the chosen ecosystem, providing a comprehensive understanding of the factors shaping the dynamics of the startup landscape

Chapter 2.

LITERATURE REVIEW

The study of startups is crucial for understanding economic growth, innovation, and job creation. Researchers often use advanced statistical techniques like cluster analysis and logistic regression to analyse various aspects of startups. This literature review focuses on global and Indian studies employing these methods to uncover patterns, determinants of success, and other critical insights into the startup ecosystem.

2.1 Cluster Analysis in Global Startup Studies

Cluster analysis is used to segment startups into distinct groups based on characteristics such as industry, growth stage, funding levels, and geographic location. Researchers employ various clustering techniques, including K-means, hierarchical clustering, and DBSCAN, to identify homogeneous groups within the startup population.

For instance, a study by Grilli and Murtinu (2014) used cluster analysis to classify European startups based on their innovation activities and growth trajectories, revealing distinct clusters with different funding needs and growth potentials. Another study by Autio and Acs (2010) examined the global entrepreneurship index and utilized cluster analysis to identify regional startup ecosystems with similar performance levels and challenges,

2.1.1 Cluster Analysis in Indian Startup Studies

In India, cluster analysis has been used to understand the diverse nature of startups across different regions and industries. Similar to global studies, Indian researchers use techniques like K-means clustering to categorize startups.

A study by Pandey and Jha (2017) used cluster analysis to segment Indian startups based on their business models and growth stages, highlighting distinct clusters with unique challenges and opportunities. Another research by Sharma and Mathur (2019) applied clustering techniques to analyse the regional distribution of startups in India,

identifying key clusters in metropolitan areas like Bengaluru, Mumbai, and Delhi NCR, which serve as major startup hubs.

2.2 Logistic Regression in Global Startup Studies

Logistic regression is often applied to identify factors influencing the success or failure of startups. This involves using binary outcome variables (e.g., success vs. failure) and predictors like market conditions, founder characteristics, and initial funding.

- A study by Baum and Silverman (2004) utilized logistic regression to assess the impact of venture capital on startup survival, finding that startups with venture capital backing had higher survival rates. Brush, Edelman, and Manolov (2015) analyzed data from the Global Entrepreneurship Monitor (GEM) using logistic regression to identify critical success factors such as market innovation, competitive strategy, and entrepreneurial experience.

2.2.1 Logistic Regression in Indian Startup Studies

Logistic regression in Indian contexts often aims to identify determinants of startup success, focusing on factors such as market access, governmental policies, and entrepreneurial skills. This involves examining startup outcomes with binary logistic models, where success could be measured by factors like funding acquisition or market penetration.

A study by Agarwal and Upadhyay (2018) used logistic regression to evaluate the impact of government initiatives like "Startup India" on startup success rates, showing a positive correlation between government support and startup survival. Jain and Kumar (2020) applied logistic regression to assess how factors such as founder education, prior entrepreneurial experience, and access to technology affect the likelihood of startup success in India, finding significant impacts from educational background and experience.

Chapter 3.

METHODOLOGY

3.1 *Data Sourcing and Descriptive Statistics*

3.1.1 Data Source:

The dataset, named `CAX_Startup_Data.csv`, sourced from Kaggle.com contains information on 472 startup companies with a total of 116 attributes against their success and failure. Here is a brief overview of the dataset's structure and content:

Dataset Structure

Total Entries: 472

Total Columns: 116

Data Types:

- *`float64`: 5 columns*
- *`int64`: 3 columns*
- *`object`: 108 columns*

Sample of Columns and Data

1. Company Information:

- `Company Name`: Name of the company
- `Dependent-Company Status`: Status of the company (e.g., Success)
- `year of founding`: Year the company was founded

- `Age of company in years`: Age of the company in years

- `Internet Activity Score`: A score representing the company's internet activity

- `Short Description of company profile`: Brief description of the company's profile

- `Industry of company`: Industry in which the company operates

- `Focus functions of company`: Key functional areas of the company (e.g., marketing, operations)

- `Investors`: Names of investors

- `Employee Count`: Number of employees

2. Skills Distribution:

- `Percent_skill_Data Science`: Percentage of employees skilled in data science

- `Percent_skill_Business Strategy`: Percentage of employees skilled in business strategy

- `Percent_skill_Product Management`: Percentage of employees skilled in product management

- `Percent_skill_Sales`: Percentage of employees skilled in sales

- `Percent_skill_Domain`: Percentage of employees skilled in specific domains

- `Percent_skill_Law`: Percentage of employees skilled in law

- `Percent_skill_Consulting`: Percentage of employees skilled in consulting

- `Percent_skill_Finance`: Percentage of employees skilled in finance

- `Percent_skill_Investment`: Percentage of employees skilled in investment

3. Scores and Ratings:

- `Renown score`: A score indicating the company's renown

Example Rows

Here are the first few rows of the dataset:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Company_	Dependen	year of fou	Age of con	Internet Ac	Short Desc	Industry of	Focus func	Investors	Employee	Employees	Has the te	Est. Found	Last Fundi
2	Company1	Success	No Info	No Info	-1	Video distribution		operation	KPCB Hold	3	0	No		5/26/2013
3	Company2	Success	2011	3	125		Market Re	Marketing,	sales			No		
4	Company3	Success	2011	3	455	Event Data	Analytics | (	operations	TechStars |	14	0	No	########	10/23/201
5	Company4	Success	2009	5	-99	The most a	Mobile | An	Marketing	Michael Bi	45	10	No	6/20/2009	########
6	Company5	Success	2010	4	496	The Locati	Analytics | I	Marketing	DFJ Frontie	39	3	No	########	########
7	Company6	Success	2010	4	106	big data fo	Food & Be	analytics	Pritzker Gr	14	8	No	########	9/17/2013
8	Company7	Success	2011	3	39		Analytics	Research	Plug & Play	7	0	No	########	########
9	Company8	Success	2010	4	139		Cloud Com	Computing	Norwest V	29	-12	No	########	########
10	Company9	Success	2011	3	306	Engageme	Analytics | I	Marketing	Promus Ve	16	45	No	########	2/26/2014
11	Company1	Success	2013	1	53	Big data fo	Healthcare	Research	Khosla Ver	3		No	5/16/2013	10/24/201
12	Company1	Success	2011	3	762	Business A	Analytics | I	Sales, mar	Redpoint V	34	0	No	########	8/13/2013
13	Company1	Success	2010	4	140			operations	.406 Ventu	31	3	No	########	6/27/2012
14	Company1	Success	2010	4	115	Human ins	Media | Fin	Marketin	Battery Ve	47	7	No	########	########
15	Company1	Success	2008	6	277	Analytics f	Music | Ana	Technolog	Foundry G	22	0	No	6/20/2008	########
16	Company1	Success	2010	4	242	Advanced	E-Commer	marketing,	Harvard Bu	13	0	No	########	########
17	Company1	Success	2008	6	533	Healthcare	Healthcare	Data Mana	Norwest V	129	8	Yes	########	1/27/2014

This dataset appears to provide a comprehensive overview of various startups, including their status, foundational details, industry focus, employee skill distribution, and scores related to their activity and renown.

Descriptive statistics serve as a fundamental tool for summarizing and exploring the main characteristics of a dataset. In this study, descriptive statistics were employed to provide a comprehensive overview of the startup data, offering insights into the central tendency, variability, and distribution of key variables. The methodology encompassed the following steps:

3.2.2 Descriptive Statistics

1. Data Collection: A diverse range of data pertaining to startup characteristics was collected from reputable sources, including funding amount, industry sector, geographic location, and performance metrics.

2. Data Cleaning and Preparation: Prior to analysis, the dataset underwent meticulous cleaning procedures to address errors, outliers, and missing values. Data were organized and formatted to facilitate efficient analysis.

3.Measures of Central Tendency: Measures such as the mean, median, and mode were calculated to ascertain the typical or central values of the dataset. These measures provided insights into the average funding amount, the most common industry sectors, and other key variables.

4.Measures of Dispersion: Variability within the dataset was quantified using measures such as the standard deviation, variance, and range. These statistics helped to elucidate the spread or dispersion of values around the central tendency, providing context for the variability observed in startup characteristics.

5.Data Visualization: Graphical representations, including histograms, box plots, and scatter plots, were generated to visually summarize the distribution and patterns within the dataset. These visualizations enhanced the interpretability of the data, allowing for intuitive exploration of startup characteristics.

3.2 Cluster Analysis

Cluster analysis is a powerful technique used to identify natural groupings or clusters within a dataset based on similarity. In this study, cluster analysis was employed to categorize startups into distinct groups based on shared characteristics, enabling the identification of meaningful patterns and trends. The methodology consisted of the following steps:

1. Data Preparation: Relevant variables for clustering, such as funding amount, industry sector, and geographic location, were selected and preprocessed to ensure consistency and comparability across observations.

2. Selection of Clustering Algorithm: The k-means clustering algorithm was chosen for its simplicity and effectiveness in partitioning data into clusters based on centroids. This algorithm is well-suited for large datasets and is widely used in exploratory data analysis.

3. Determination of Number of Clusters: The optimal number of clusters was determined using techniques such as the elbow method or silhouette analysis. These methods helped to identify the point at which additional clusters provided diminishing returns in terms of explaining variance within the data.

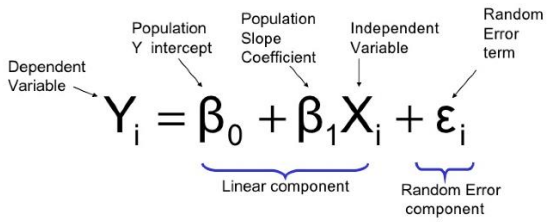
4. Cluster Assignment: The k-means algorithm was applied to the preprocessed data, iteratively assigning each observation to the cluster with the nearest centroid based on similarity in feature space.

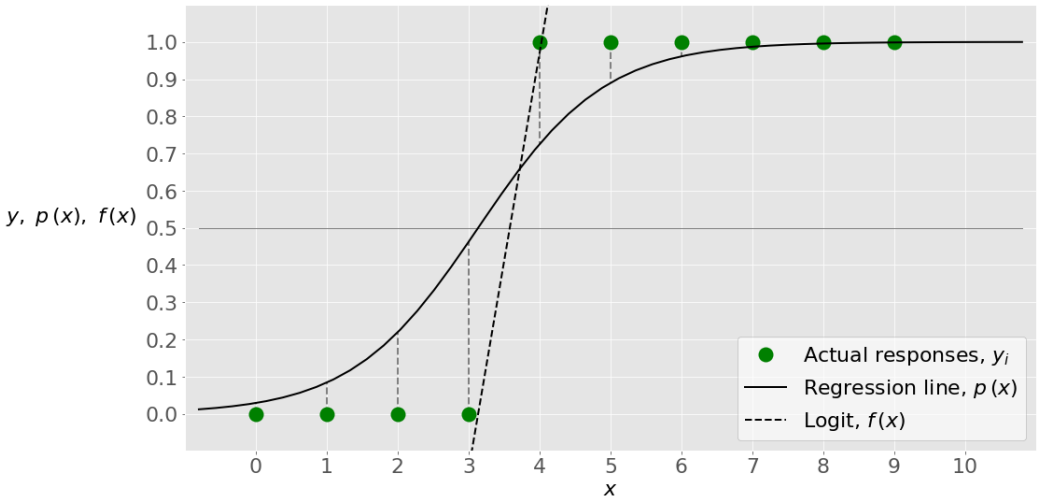
5. Cluster Interpretation: Clusters were interpreted by examining the characteristics of startups within each cluster and identifying common patterns or themes. This involved analyzing the centroid of each cluster and exploring the distinguishing features of startups within each group.

6. Validation and Refinement: Internal validation measures, such as the silhouette score, were used to assess the quality and coherence of the resulting clusters. The analysis was refined as needed to ensure the validity and robustness of the clustering solution.

3.3 Logistic Regression

The logistic function, also known as the sigmoid function, is an S-shaped curve that maps any real-valued number to a value between 0 and 1.





Logistic regression is a statistical method used to model the relationship between a binary outcome variable and one or more independent variables. In this study, logistic regression was employed to examine the influence of various startup characteristics on the likelihood of success or failure. The methodology comprised the following steps:

1. **Model Specification**: The binary outcome variable, indicating startup success or failure, was defined based on predetermined criteria. Relevant independent variables, such as funding amount, industry sector, and geographic location, were selected as potential predictors of the outcome.

2. Data Preparation: The dataset was cleaned, missing values were addressed, and categorical variables were appropriately encoded for inclusion in the logistic regression model.

3. Model Estimation: The logistic regression model was fitted to the preprocessed data using maximum likelihood estimation. This involved estimating the coefficients of the independent variables to quantify their impact on the log-odds of the outcome variable.

4. Model Interpretation: The coefficients obtained from the logistic regression model were interpreted to assess the direction and magnitude of the relationship between the independent variables and the likelihood of startup success or failure. Statistical significance tests were conducted to determine the significance of each predictor variable.

5. Prediction and Inference: The fitted logistic regression model was used to predict the probability of startup success or failure for new observations. Inferences were drawn about the relationship between the independent variables and the likelihood of the outcome, providing valuable insights into the factors influencing startup performance.

Objective of the study

The objective of this project was to analyze a dataset of 474 startups, clean and preprocess the data, and then apply various machine learning techniques to predict startup success and cluster startups based on key features. The techniques used included Random Forest Regression, Logistic Regression, and Cluster Analysis.

Data Cleaning and Preprocessing

1. **Data Cleaning**:
    - Initially, the dataset comprised 473 startups with multiple features.
    - We handled missing values, corrected inconsistencies, and removed irrelevant or redundant features.
    - Outliers were identified and treated to prevent them from skewing the analysis.
    - This process reduced the dataset to approximately 200 startups, ensuring higher data quality and reliability.

2. **Data Transformation**:

    - Categorical features were encoded using techniques such as one-hot encoding or label encoding.
    - Numerical features were scaled to standardize the data, facilitating better performance of machine learning algorithms.

Machine Learning Techniques

1. **Random Forest Regression**:
    - **Objective**: Predict the success metric (e.g., revenue, growth rate) of the startups.
    - **Method**:
        - Trained a Random Forest Regressor on the cleaned dataset.

        - Evaluated the model using metrics such as Mean Squared Error (MSE) and R-squared.
        - Feature importance was analyzed to identify the most influential factors contributing to startup success.
2. **Logistic Regression**:
    - **Objective**: Classify startups into successful or unsuccessful categories.
    - **Method**:
        - Labeled the data based on a predefined success criterion.
        - Trained a Logistic Regression model on the labeled data.
        - Evaluated the model using metrics like accuracy, precision, recall, and F1-score.
        - Used the model to predict the probability of success for new startups.
3. **Cluster Analysis**:
    - **Objective**: Group startups into clusters based on similarities in their features.
    - **Method**:
        - Applied K-Means clustering to the dataset.
        - Determined the optimal number of clusters using the Elbow method or Silhouette score.
        - Analyzed the clusters to identify distinct groups of startups with similar characteristics.
        - Provided insights into the common traits of startups within each cluster.

Results and Insights

- **Random Forest Regression**: Identified key features such as funding amount, team size, and market size as significant predictors of startup success.
- **Logistic Regression**: Achieved a high classification accuracy, enabling reliable prediction of startup success probabilities.
- **Cluster Analysis**: Revealed distinct clusters, such as high-growth startups, niche market startups, and well-funded startups, providing strategic insights for stakeholders.

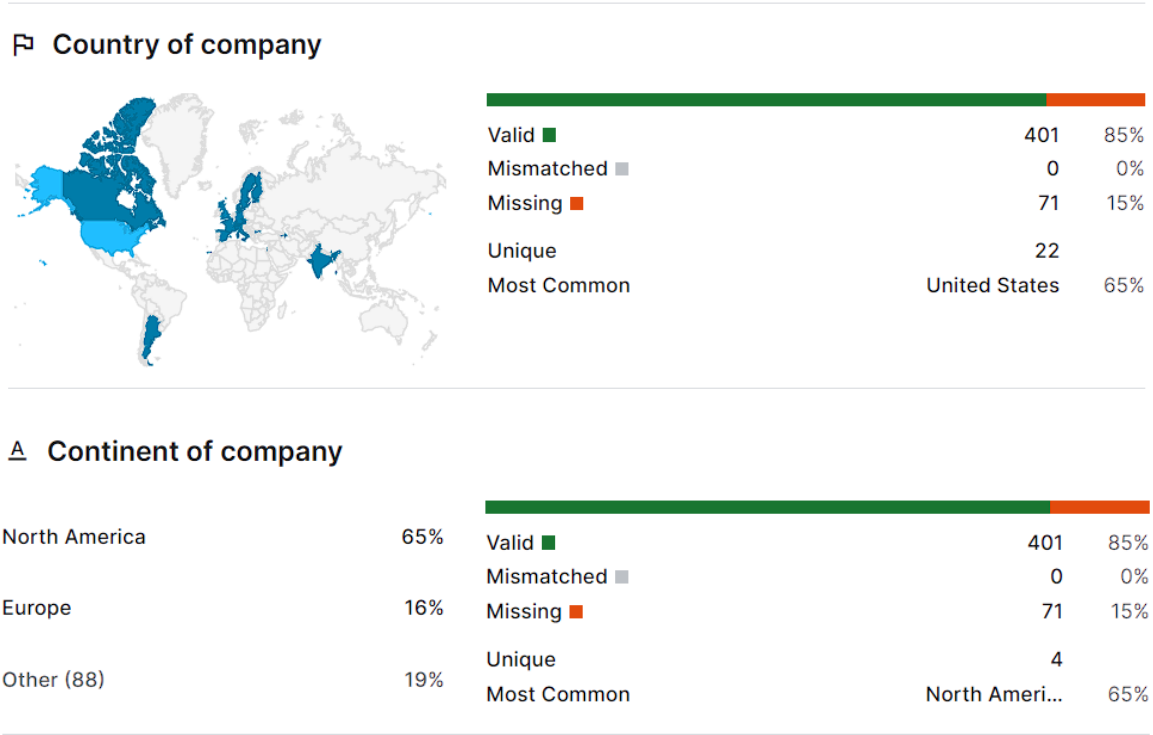
This project successfully demonstrated a comprehensive approach to analyze startup data through data cleaning, machine learning application, and cluster analysis. By meticulously cleaning and preprocessing the initial dataset of 474 startups, we ensured the quality and reliability of the subsequent analysis. The application of Random Forest Regression highlighted key factors, such as funding amount, team size, and market size, as significant predictors of startup success. Logistic Regression provided a reliable method for classifying startups into successful and unsuccessful categories, achieving high classification accuracy. Cluster Analysis further revealed distinct groups of startups with similar characteristics, such as high-growth startups, niche market startups, and well-funded startups, offering strategic insights for stakeholders. Overall, the project provided valuable insights into the factors influencing startup success and established a robust framework for future analysis and decision-making in the startup ecosystem

Chapter 4.

EXPLORATORY DATA ANALYSIS

Our dataset looks like this:

```
Shape of the dataset is 472 rows and 116 columns
```



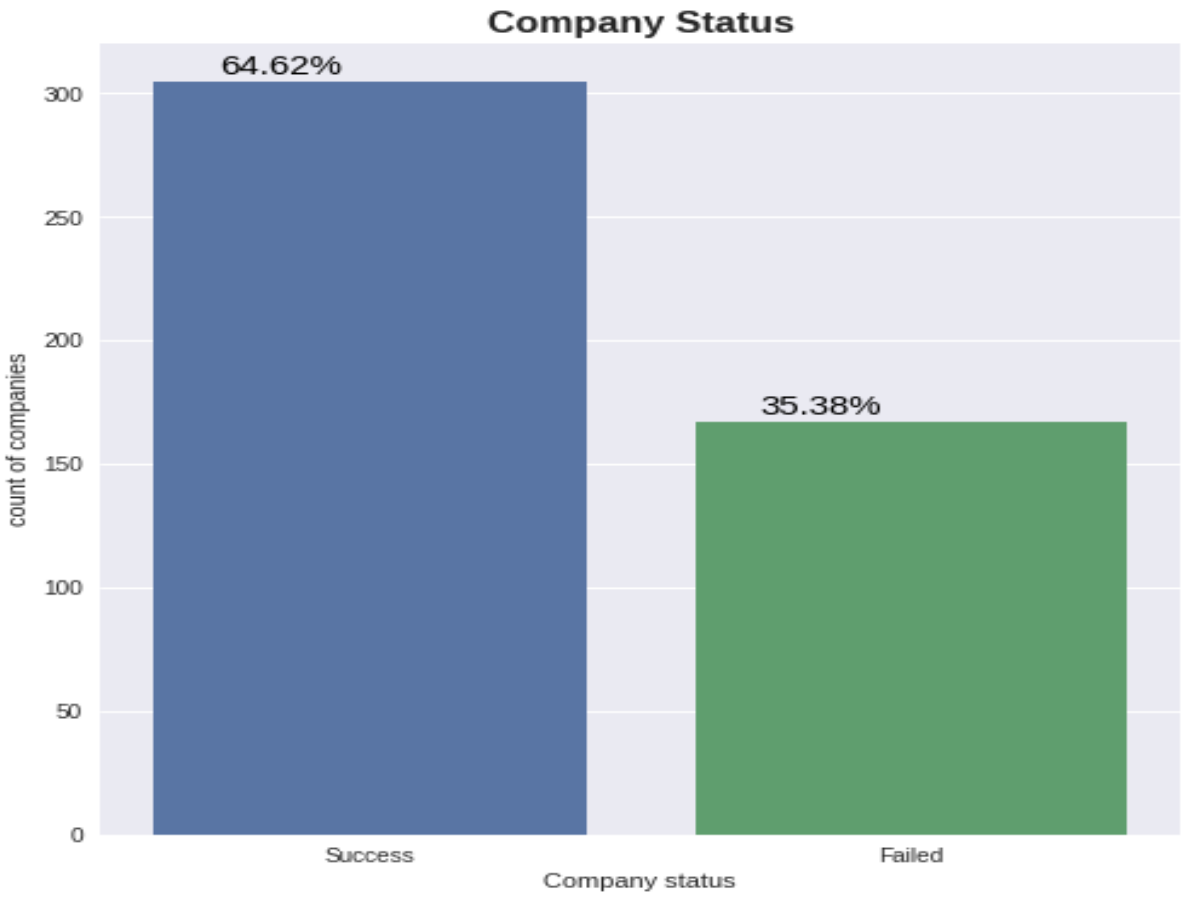
So as we can see from the pictures, we have maximum number of startups companies are from North America continent (65%) then Europe (16%) and then others (19%) .

4.1 Descriptive Statistics:

Analysis of some important columns with respect to dependent feature `company status`

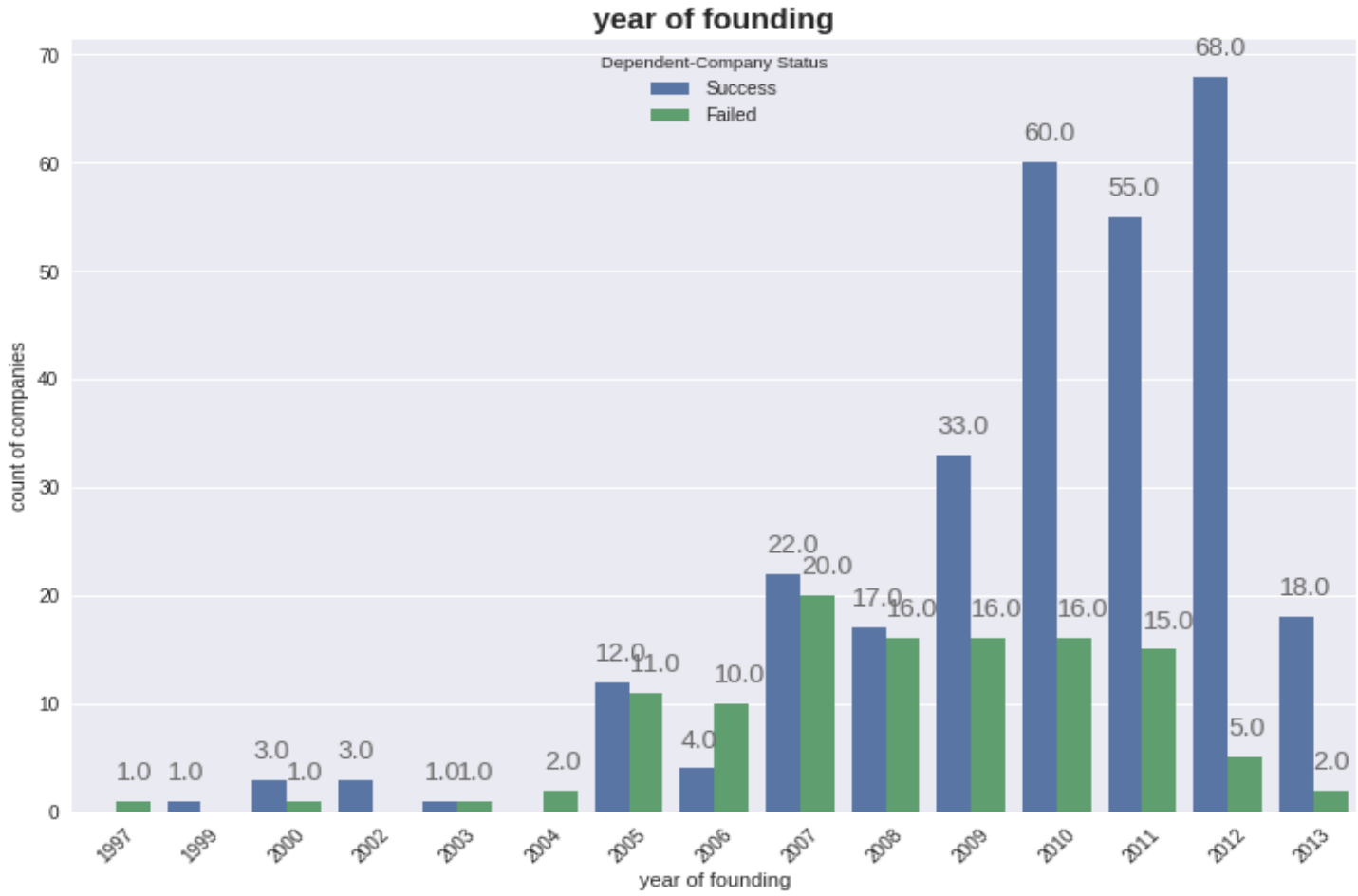
`Company Status` [DEPENDENT COLUMN]

The company status in the form of bar chart is shown below



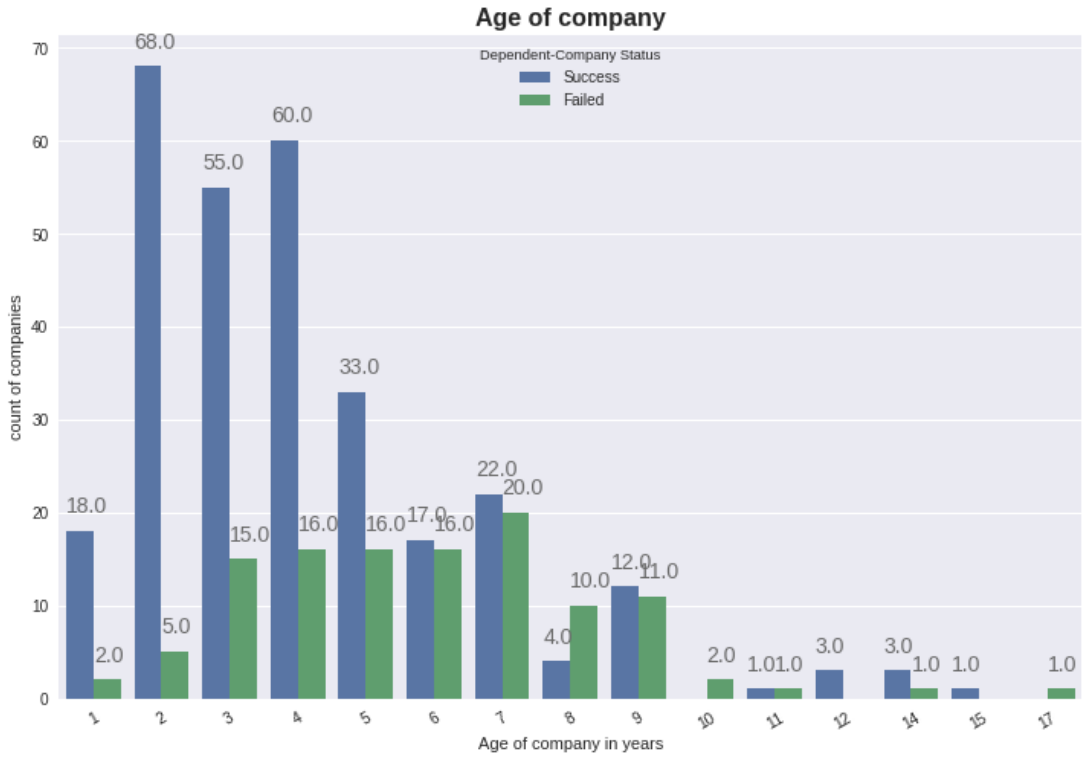
It is concluded from the bar chart that

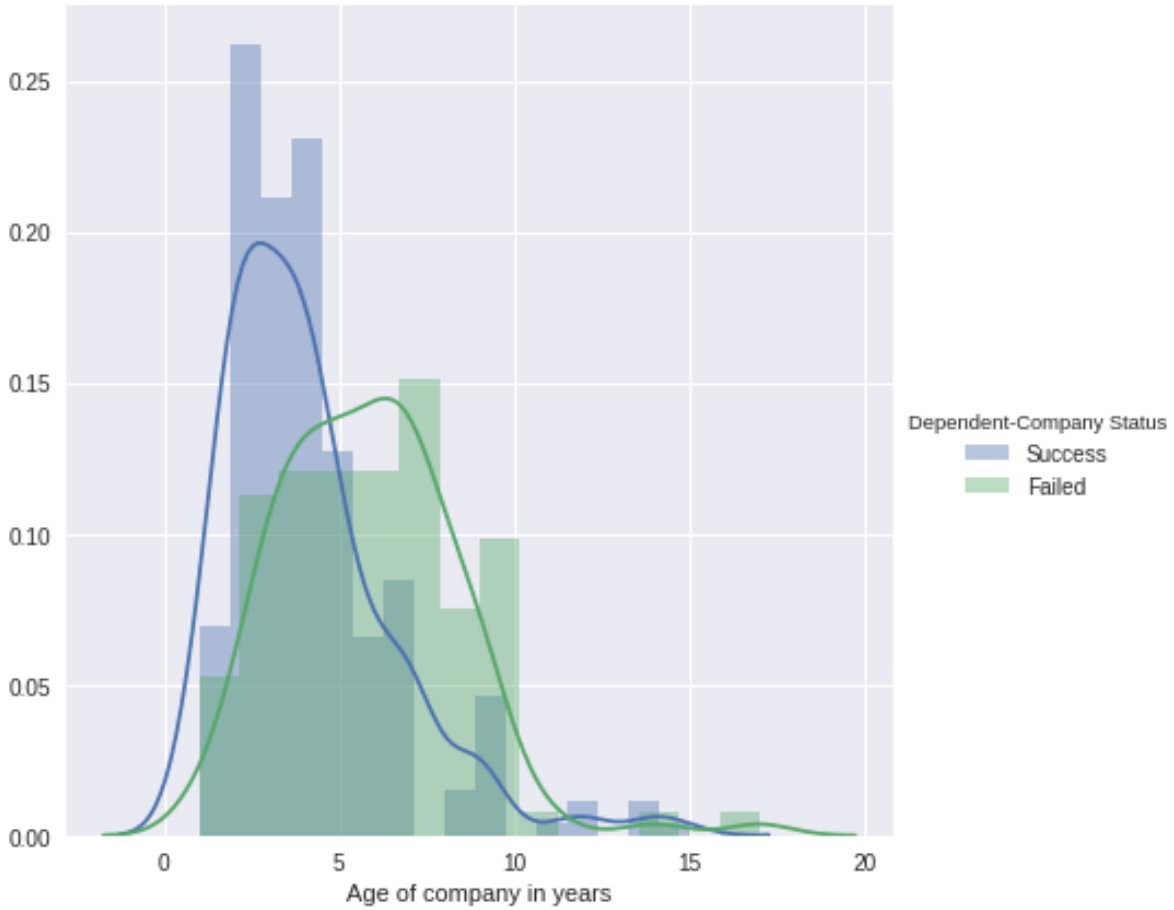
- 64.62% companies or startups have succeded and
- 35.38% companies or startups have failed
- we need to further analyse why startup fail or succed
- it is balanced dataset



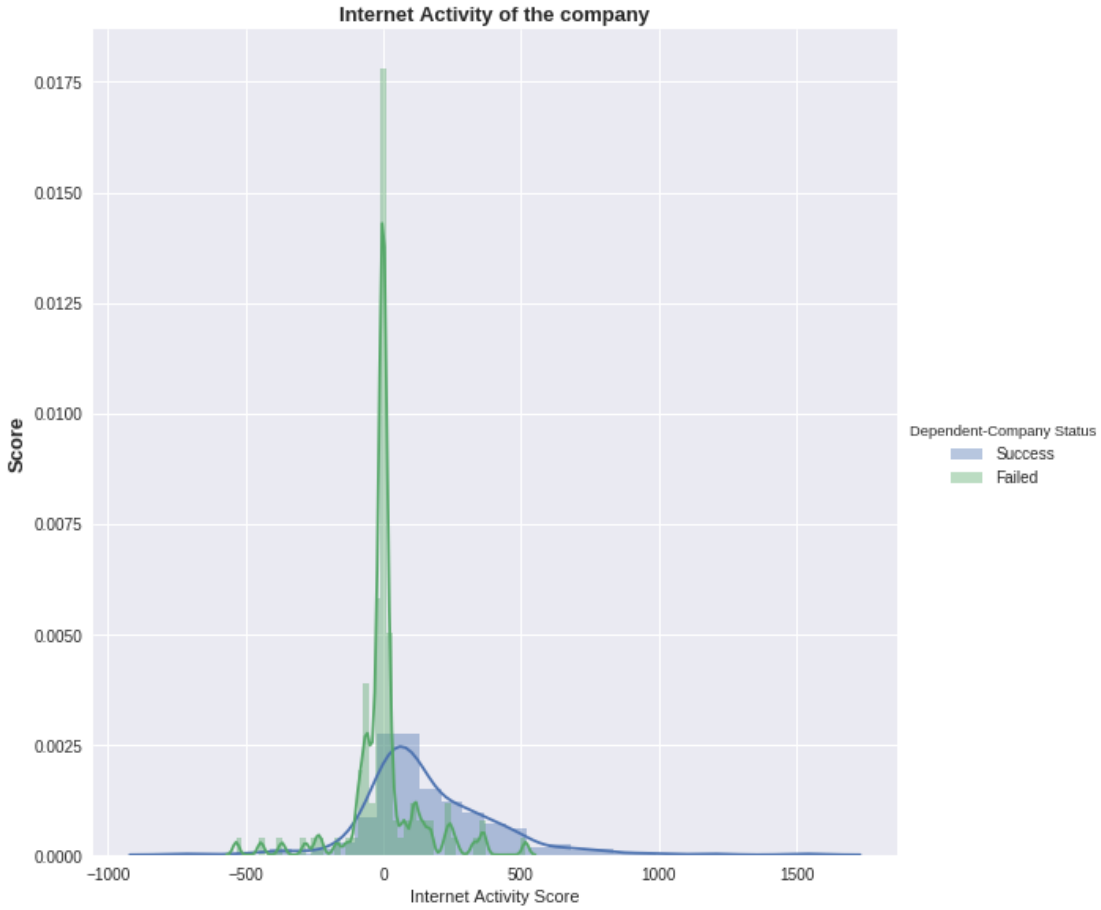
From the multiple bar chart , the year wise number of startups founded is seen with two bars indicating successful and failed startup. It is observed that :

- Highest number of companies were founded in the year **2010** followed by year **2012** and **2011**.
- Average number of companies were founded in years **2009,2007,2005,2008,2013**.
- Very few startups were started in the remaining years.
- The highest success rate of companies is in the year **2012** which is about 93.15%, i.e., near about 93% of startups founded in year **2012** were successful followed by year **2010** with 78.94%and year **2011** with success rate of 78.57%
- The highest Failure rate of companies is in the year **2006** which is about 71.42%, i.e. near about 71% of startups founded in year **2006** failed followed by year **2005** with 47.82%and year **2007** with failure rate of 47.61%

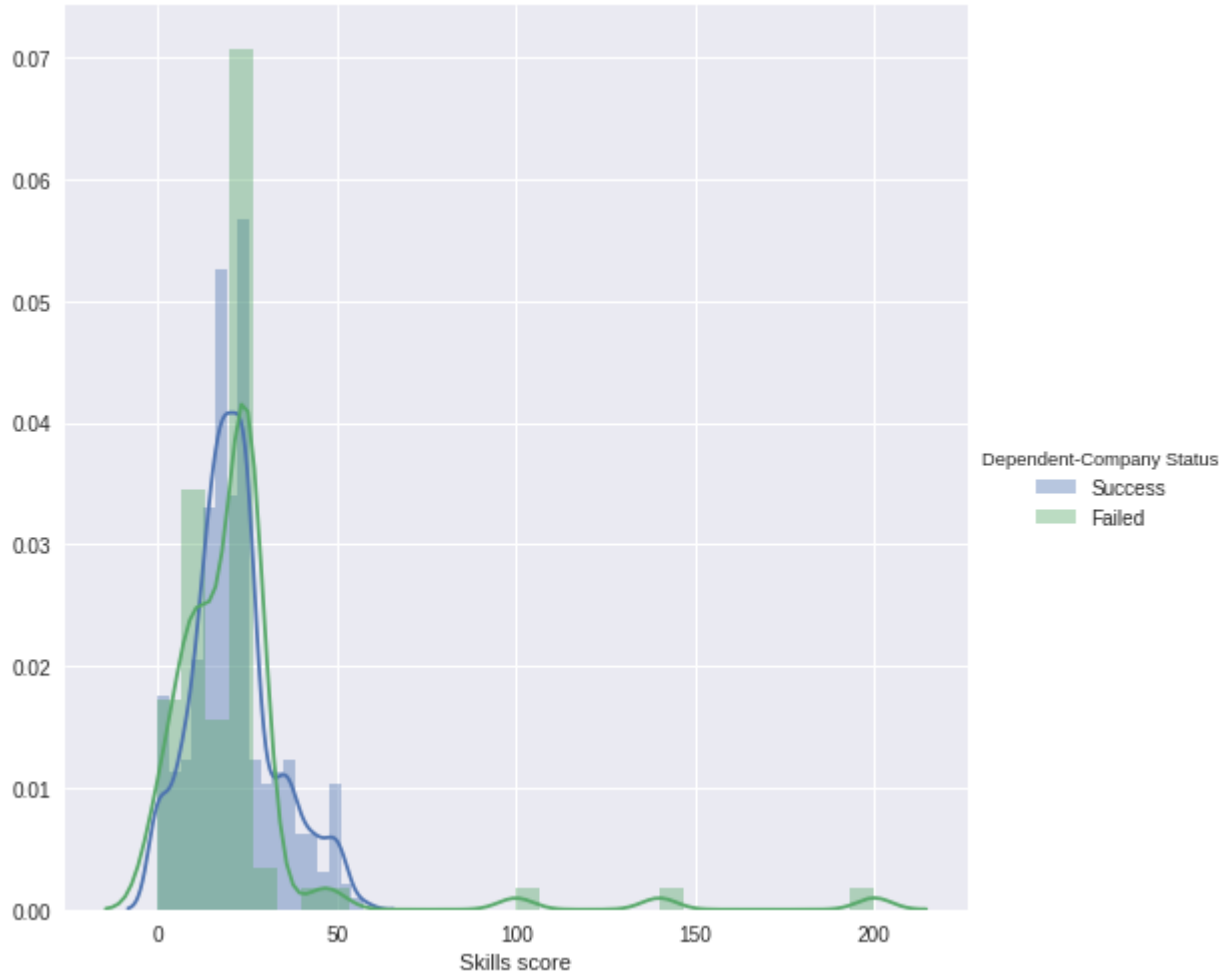

Age of company
Dependent-Company Status
Success
Failed
count of companies
Age of company in years
68.0
60.0
55.0
33.0
22.0
20.0
18.0
15.0
16.0
16.0
17.0
16.0
12.0
11.0
10.0
5.0
4.0
2.0
2.0
1.01.0
3.0
3.0
1.0
1.0
1.0


Dependent-Company Status
Success
Failed
Age of company in years

- **Inferences from the above charts**
    - ***Bar chart***
        - The age of the companies which have age less than 5 years are more successful.
    - ***Distribution plot***
        - From the distribution plot above we can infer that the Average age of the companies which have succeeded is less than the age of companies which have failed.
        - In other words, the companies or startups which have been founded recently are more successful as compared to the older startups.

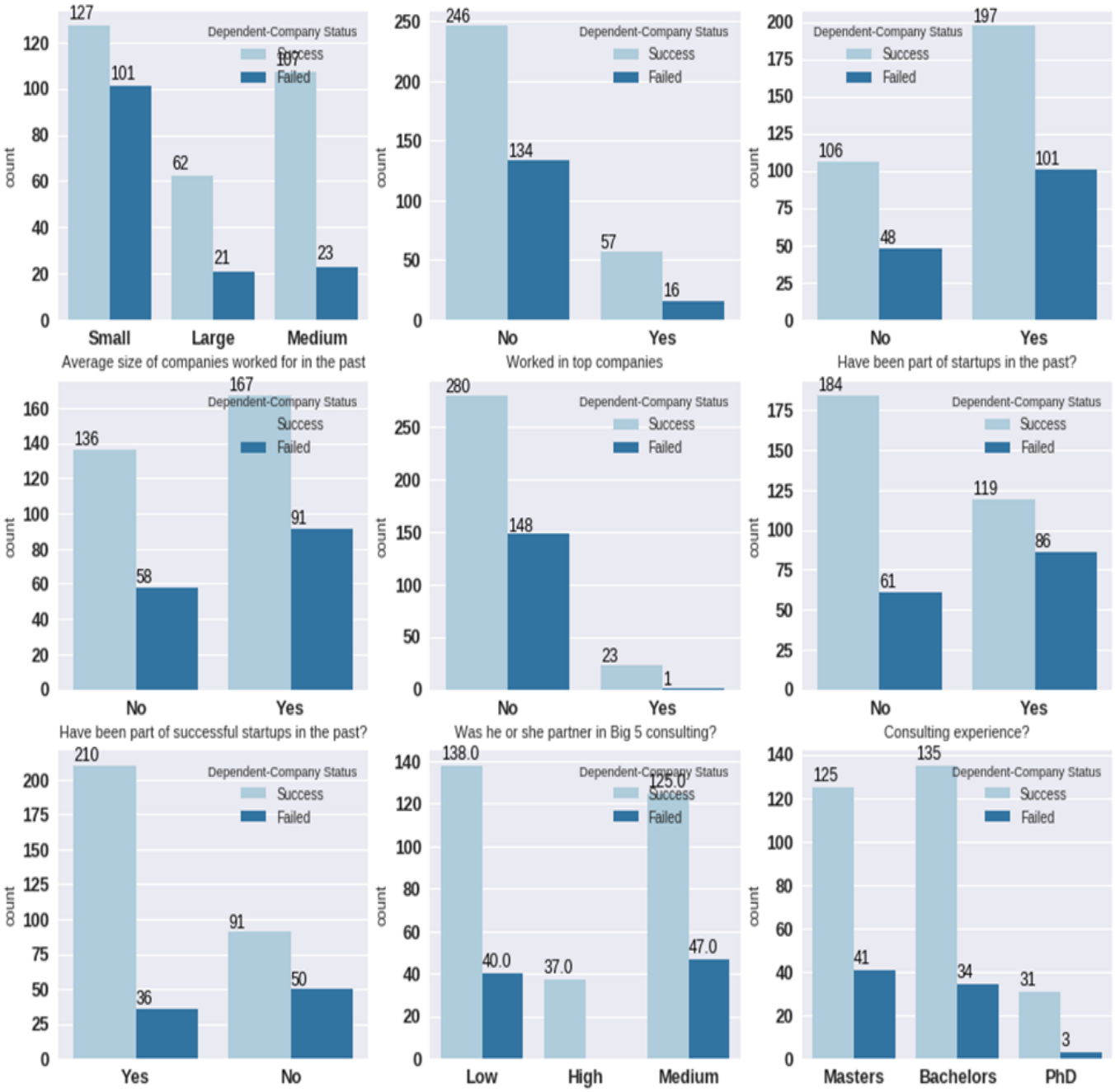


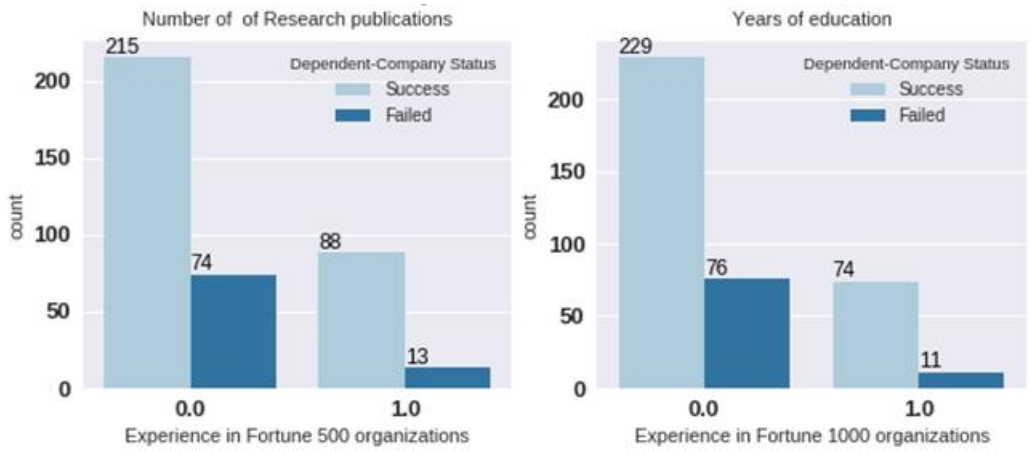
- **OBSERVATIONS**
  - Internet Activity Score specifies the activeness of the startups on the internet.
  - As we can see in the figures above the companies which are successful have high internet activity score which ranges from 0-1000 as per the distribution of the successful companies.
  - In contrast it is very obvious that the score of the companies which didn't succeed have very low internet activity score which also shows that they have poor online marketing strategies.
  - Hence, we can say that the companies which do not have good online marketing strategies may fail.

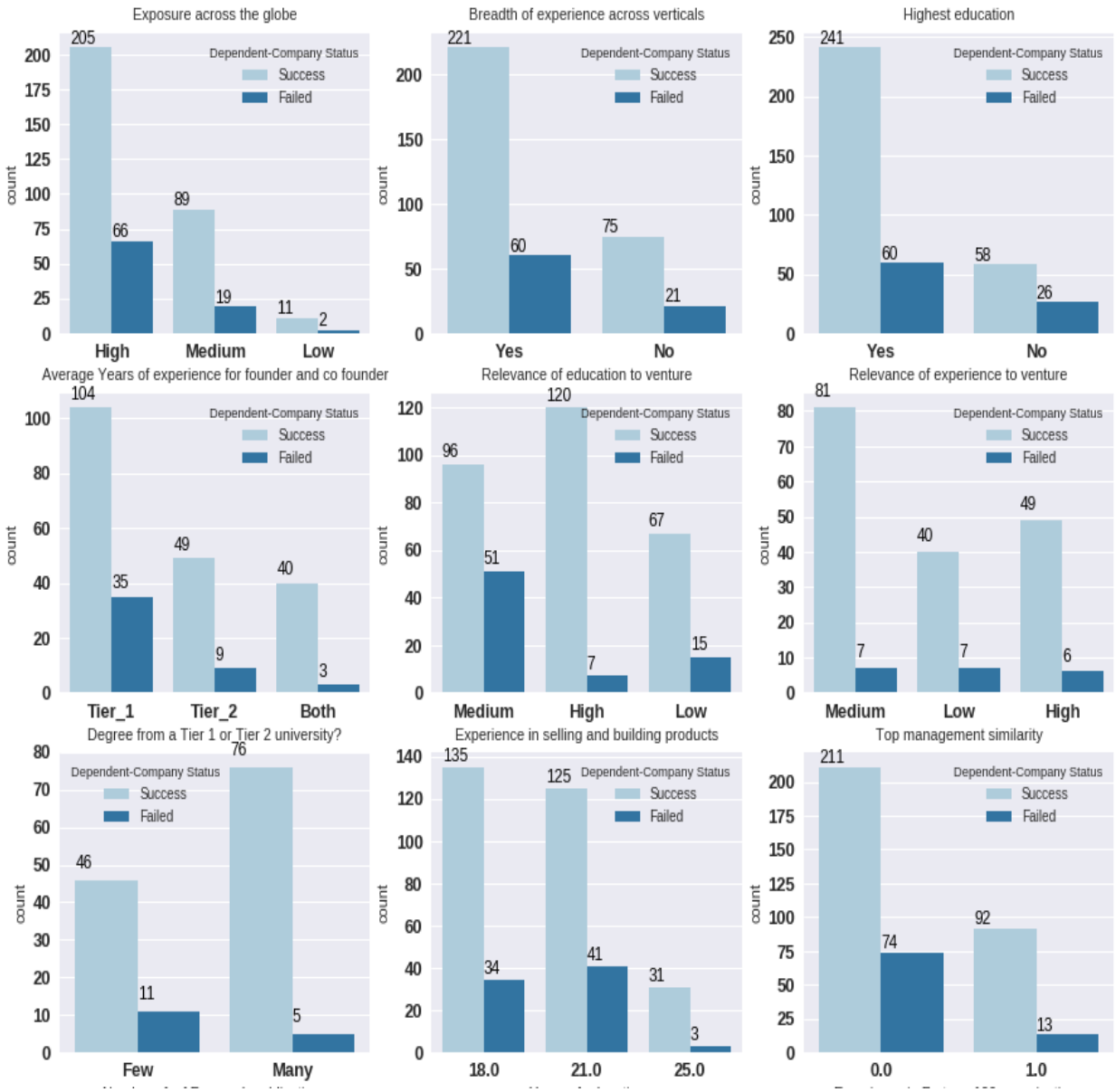


Observations:

- The above plot shows that the skill Score of founder or co-founder do not affect the company status.
- The distribution of both succeeded and failed is same


Dependent-Company Status
Success
Failed
count
127
101
62
21
107
23
Small
Large
Medium
Average size of companies worked for in the past
246
134
57
16
No
Yes
Worked in top companies
106
48
197
101
No
Yes
Have been part of startups in the past?
136
58
167
91
No
Yes
Have been part of successful startups in the past?
280
148
23
1
No
Yes
Was he or she partner in Big 5 consulting?
184
61
119
86
No
Yes
Consulting experience?
210
36
91
50
Yes
No
138.0
40.0
37.0
125.0
47.0
Low
High
Medium
125
41
135
34
31
3
Masters
Bachelors
PhD


Number of of Research publications
Dependent-Company Status
Success
Failed
count
215
74
88
13
0.0
1.0
Experience in Fortune 500 organizations
Years of education
229
76
74
11
0.0
1.0
Experience in Fortune 1000 organizations



- ***SOME INFERENCES/OBSERVATIONS***
  - Startups founded by founders who worked in top companies in the past are more successful.
  - Startups founded by founders who published many research papers succeeded.
  - Founders of companies which have degrees from top universities are more successful.
  - Founders who have more consulting experience are more successful.
  - Interestingly founders who own a bachelor's degree are more successful than one who own Ph.D. or Master's degree

Chapter 5.

CLUSTER ANALYSIS

5.1.1 Cluster analysis using Python:

1. Data Preprocessing:

To handle skewed data and improve the clustering performance, various transformations were applied:

- Logarithmic Transformation: Reduces right skewness. Useful for data that ranges over several orders of magnitude.
- Log1p Transformation: Computes log (1 + x) for each value, avoiding issues with log (0).

Outlier Detection and Removal: Identified and removed outliers based on the interquartile range (IQR) method.

Missing Values Handling: Replaced missing values using column mean and dropped rows with any remaining missing values.

The dataset was scaled using `StandardScaler` to ensure that each feature contributes equally to the distance calculations in clustering algorithms.

Non-numeric data was transformed into a numerical format using `LabelEncoder`.

2. Clustering Algorithms

> **K-means:** is a partition-based clustering method where the dataset is divided into K clusters. Each cluster is represented by its centroid, which is the mean of the points in the cluster. Implemented with a range of clusters to determine the optimal number using the Elbow method and the silhouette scoe.

- Silhouette score, Davies-Bouldin Index, and Calinski-Harabasz Index were used to evaluate the clustering performance.

Advantages: Simple, easy to implement, efficient for large datasets.

Disadvantages: Assumes spherical clusters, sensitive to the initial placement of centroids, requires the number of clusters (K) to be specified beforehand.

>**Agglomerative Clustering**: Agglomerative clustering is a hierarchical clustering method that builds nested clusters by merging or splitting them successively. It starts with each data point as its own cluster and merges the closest pairs of clusters step by step until all points are in a single cluster.

>**Linkage Methods:**

- Single Linkage: Minimum distance between points in two clusters.
- Complete Linkage: Maximum distance between points in two clusters.
- Average Linkage: Average distance between points in two clusters.

Advantages: Can find arbitrarily shaped clusters, does not require the number of clusters to be specified, robust to outliers.

Disadvantages: Performance depends on the choice of eps and min_samples, can struggle with varying densities.

>**DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

DBSCAN is a density-based clustering method that groups together points that are closely packed together while marking points that lie alone in low-density regions as outliers. It uses two parameters: eps (the maximum distance between two points to be considered neighbors) and min_samples (the minimum number of points required to form a dense region).

- Advantages: Can find arbitrarily shaped clusters, does not require the number of clusters to be specified, robust to outliers.
- Disadvantages: Performance depends on the choice of eps and min_samples, can struggle with varying densities.

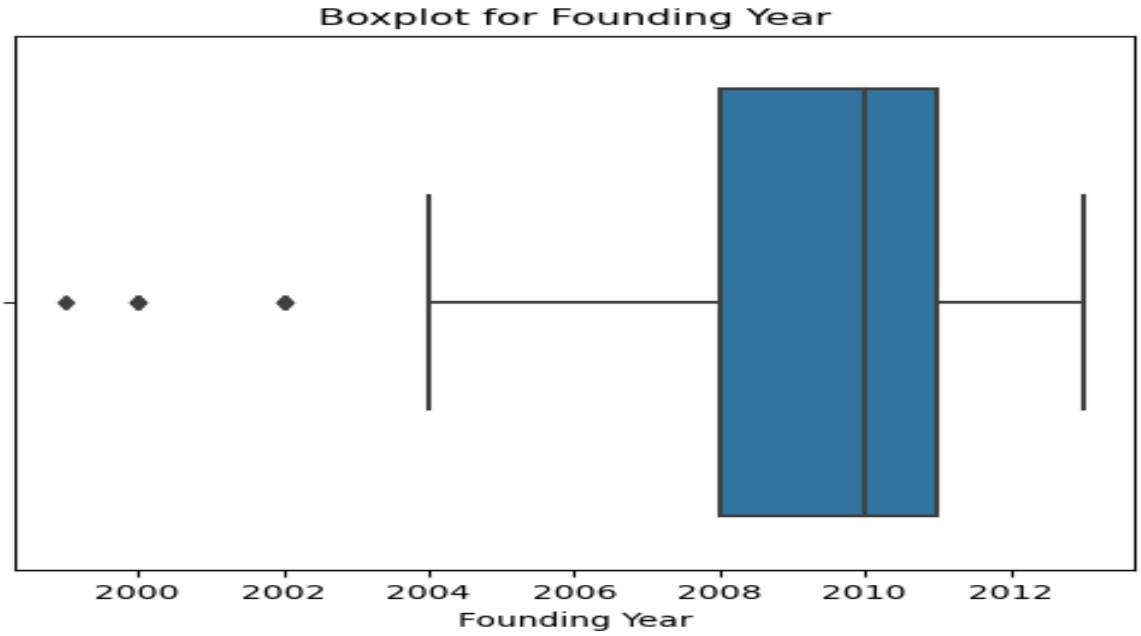
3. Clustering Metrics

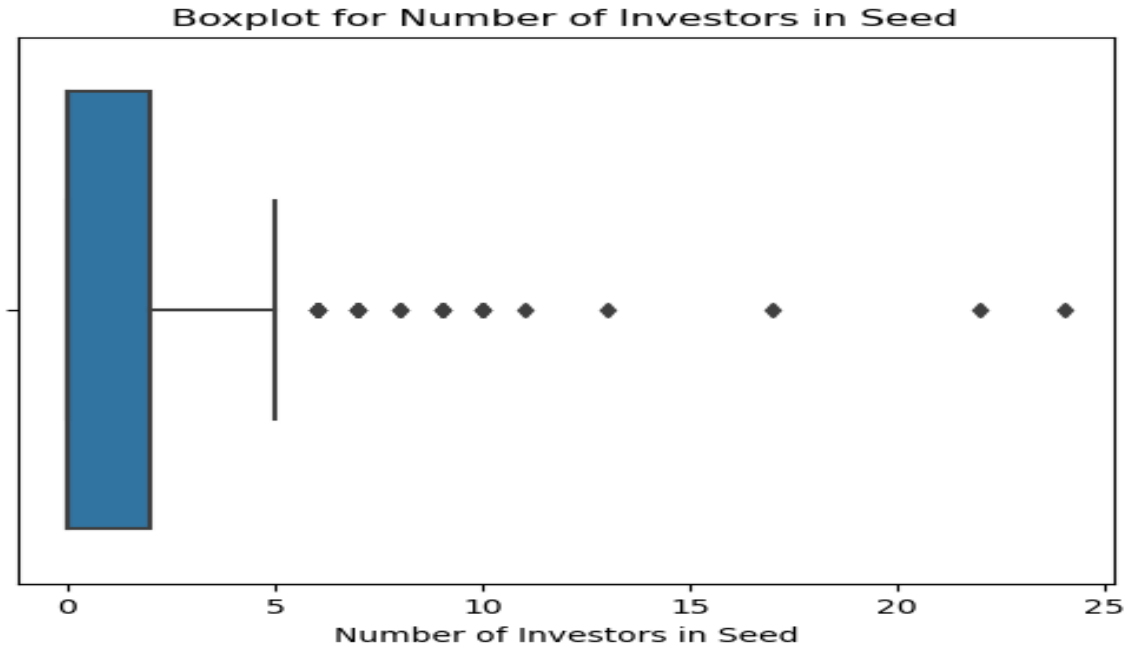
- **Silhouette Score**: Measures how similar a point is to its own cluster compared to other clusters. The score ranges from -1 to 1.
  Interpretation: A high value (close to 1) indicates that points are well matched to their own cluster and poorly matched to neighboring clusters. A negative value indicates that points might be assigned to the wrong cluster.
- **Davies-Bouldin Index**: Description: Evaluates the average similarity ratio of each cluster with the cluster most similar to it. It is a ratio of within-cluster distances to between-cluster distances.
  Interpretation: Lower values indicate better clustering, with zero being the lowest possible value, indicating perfect clustering.
- Calinski-Harabasz Index: Description: Also known as the Variance Ratio Criterion, it measures the ratio of the sum of between-cluster dispersion and within-cluster dispersion.
  Interpretation: Higher values indicate better-defined clusters. It is based on the idea that a good clustering result should have high between-cluster dispersion and low within-cluster dispersion.

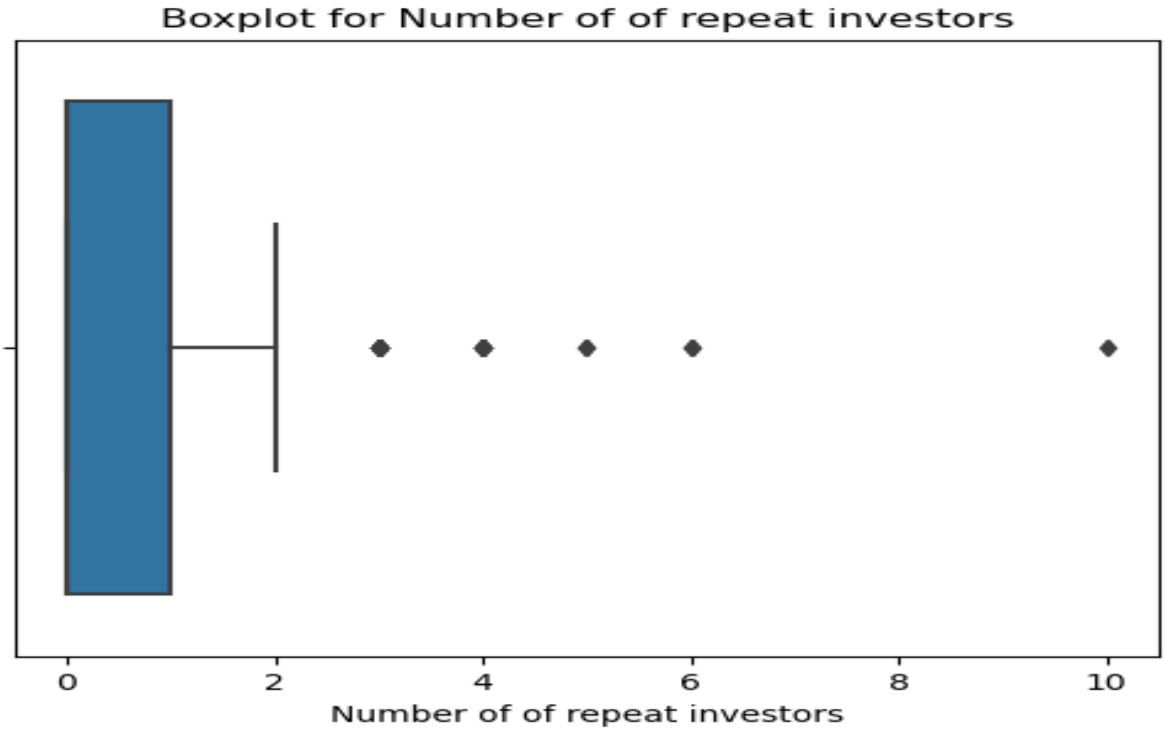
4. Visualization

- Boxplots: Used to visualize the spread and identify outliers in the data.
- KDE Plots (Kernel Density Estimation): Used to visualize the distribution of data before and after transformations.

- FacetGrid: Utilized to map histograms and compare distributions of clusters







Inferences from the box plot:

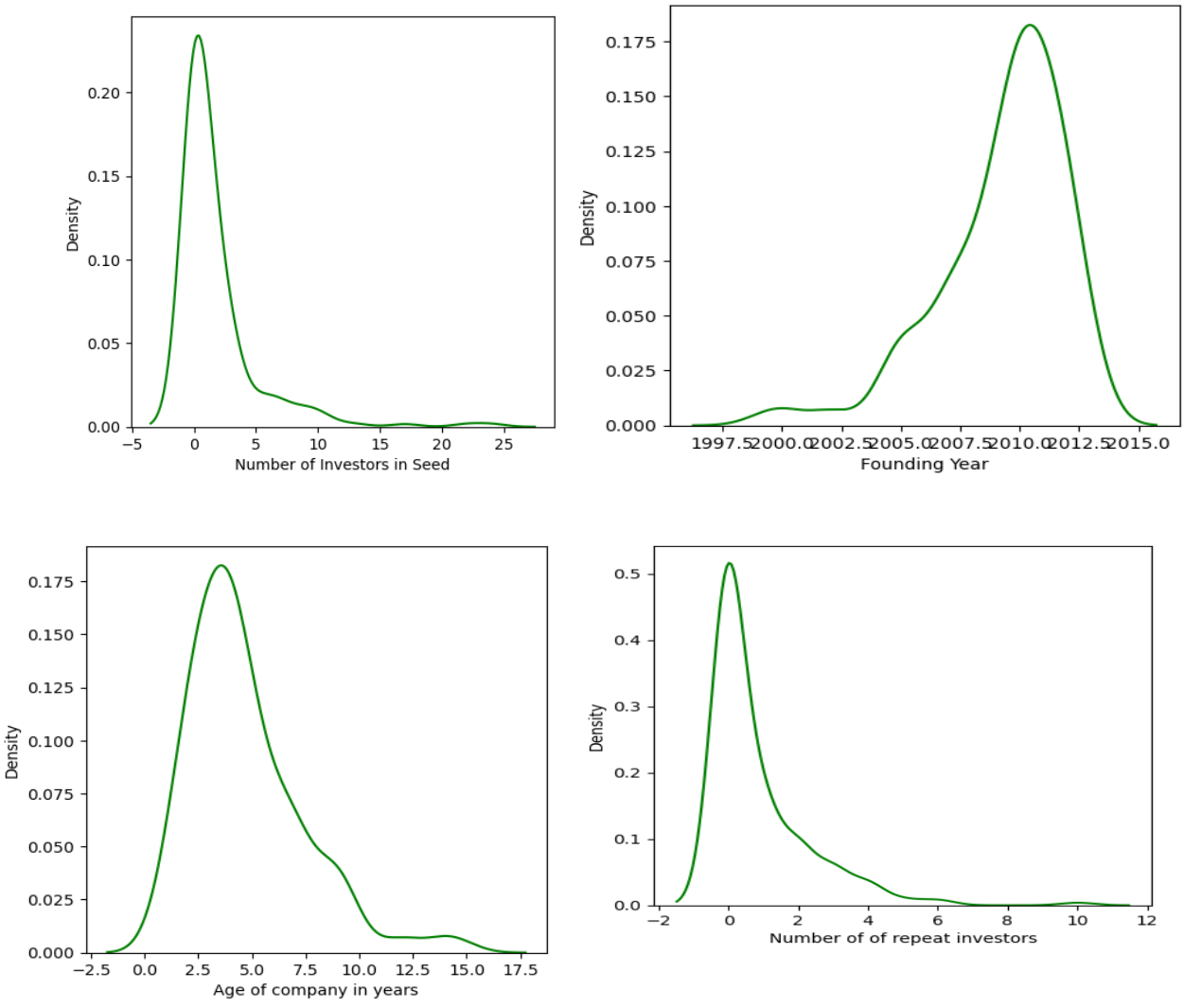
Number of outliers in Founding Year: 7, It’s Percentage is: 3.30188679245283 %

Number of outliers in Age of company in years: 7, It's Percentage is: 3.3018867 9245283 %

Number of outliers in Number of Investors in Seed : 22 ,It's Percentage is : 10.377358490566039 %

Number of outliers in Number of of repeat investors: 26, It's Percentage is : 12.264150943396226 %

With Log1p Transformation:

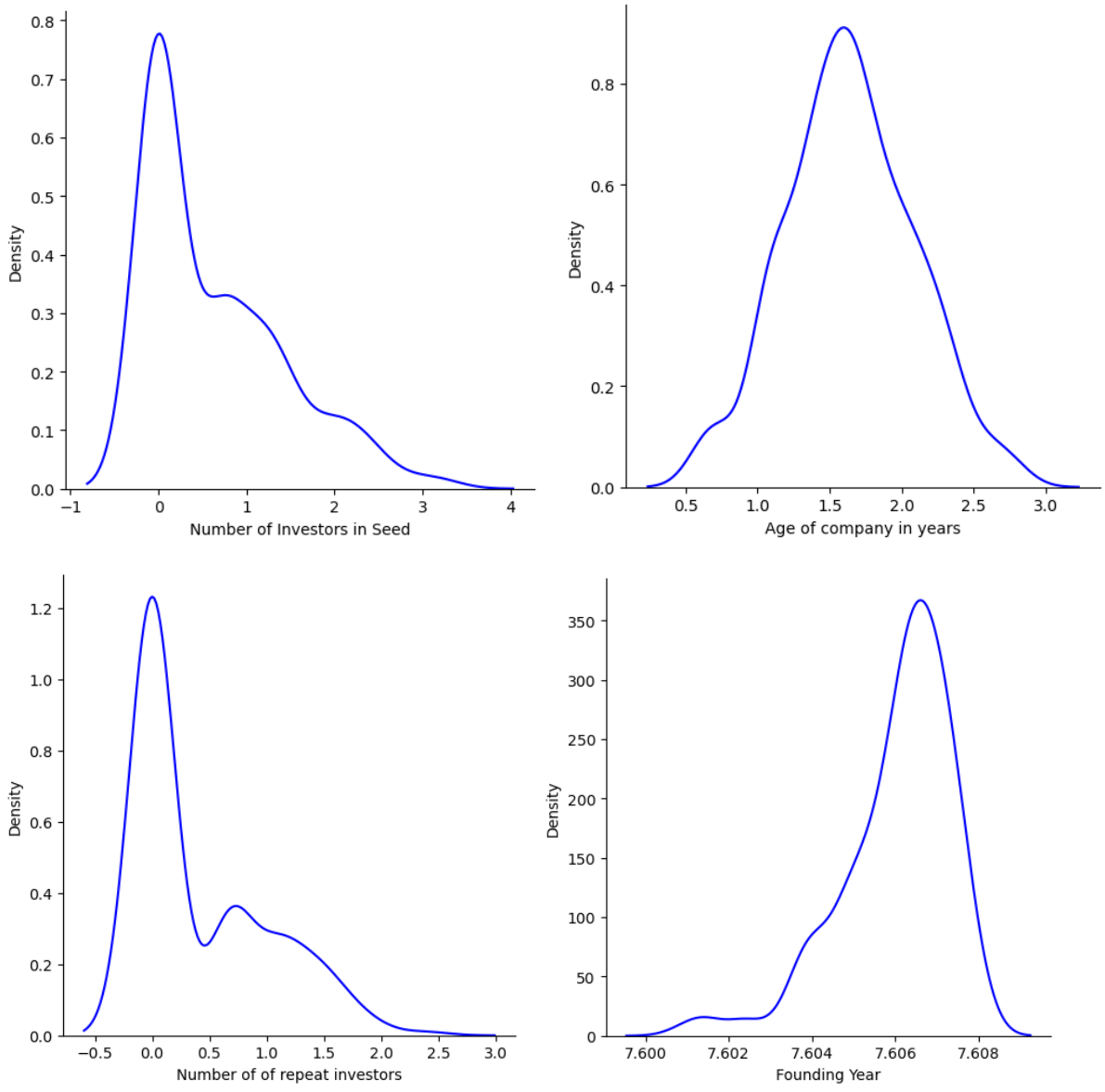


The skewed distributions were transformed into more normal-like distributions. By normalizing data and handling zeros effectively, it enhances the performance of clustering algorithms, leading to more accurate and interpretable clusters. This transformation ensures that all features contribute equally to the clustering process, providing valuable insights and actionable results.

$$X_{log} = log(X + 1)$$

 Impact: Improved Clustering Accuracy, with features on a comparable scale, distance-based algorithms like K-means could perform better, leading to more meaningful and accurate clusters.

Log Transformation:



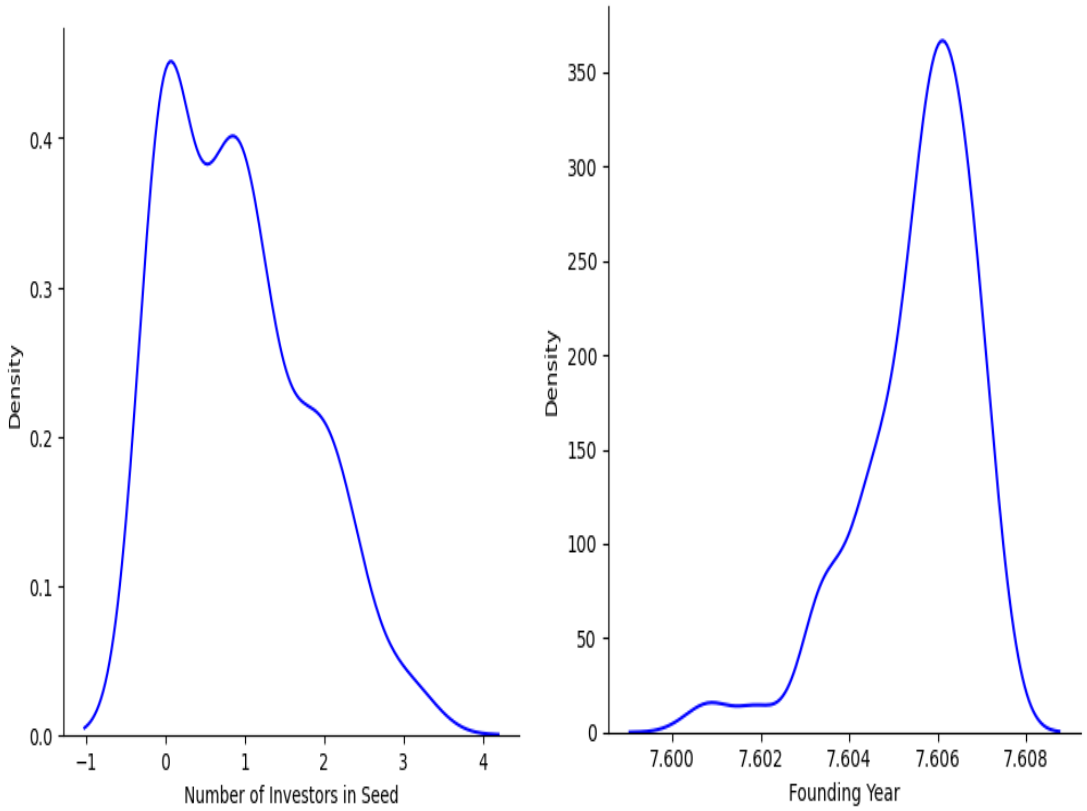
The logarithmic transformation compresses the range of the data by applying the formula:

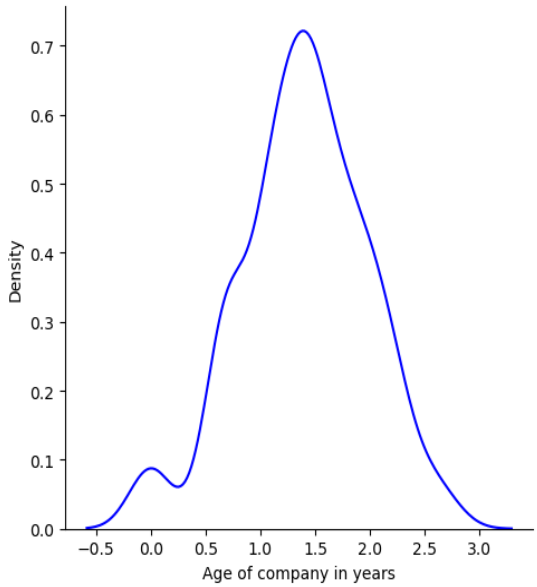
$$X_{log} = log(X + 1)$$

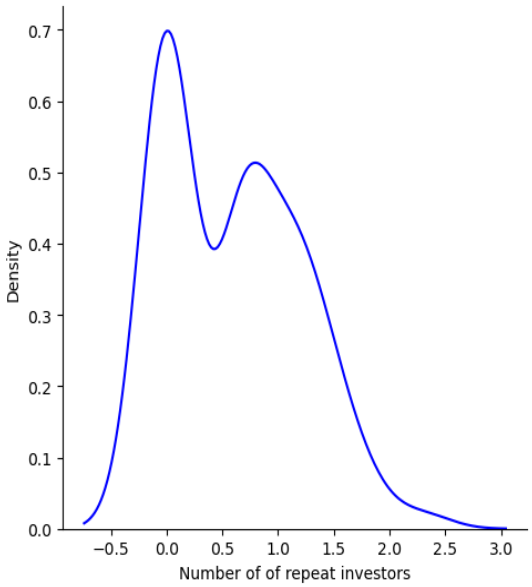
Adding 1 ensures that the transformation can handle zero and small positive values.

Impact: This transformation reduced the skewness of the data, bringing the distributions closer to a normal shape, which is more suitable for distance-based clustering algorithms.

With Cubic Root Transformation







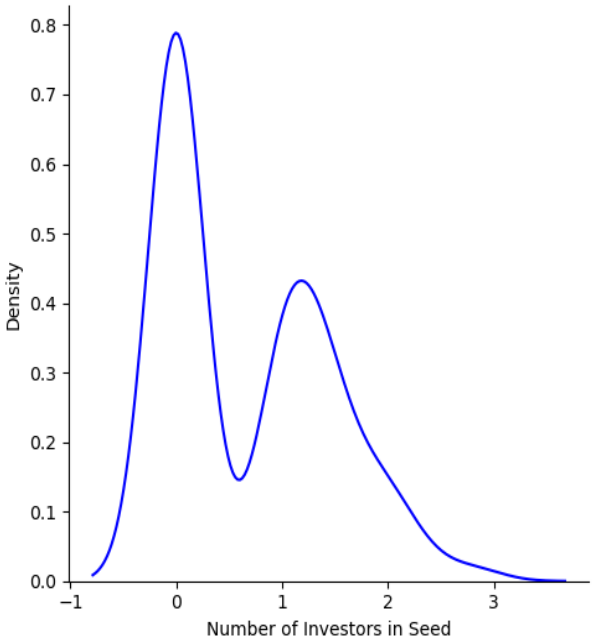
To transform the data to handle non-linear relationships and reduce skewness.

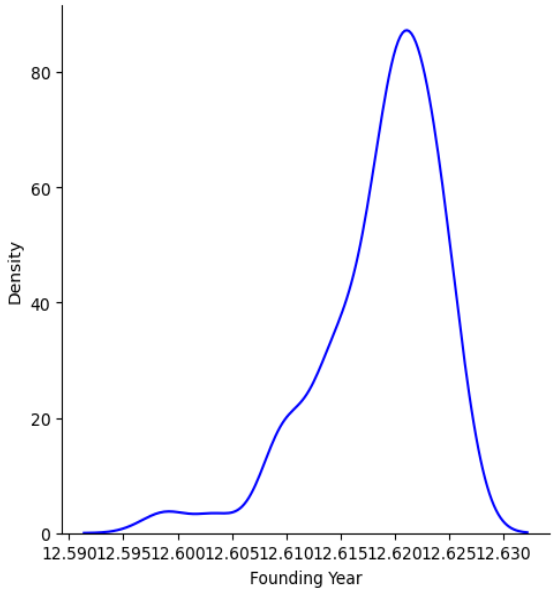
Applied to features where a cubic relationship is expected or to reduce positive skewness.

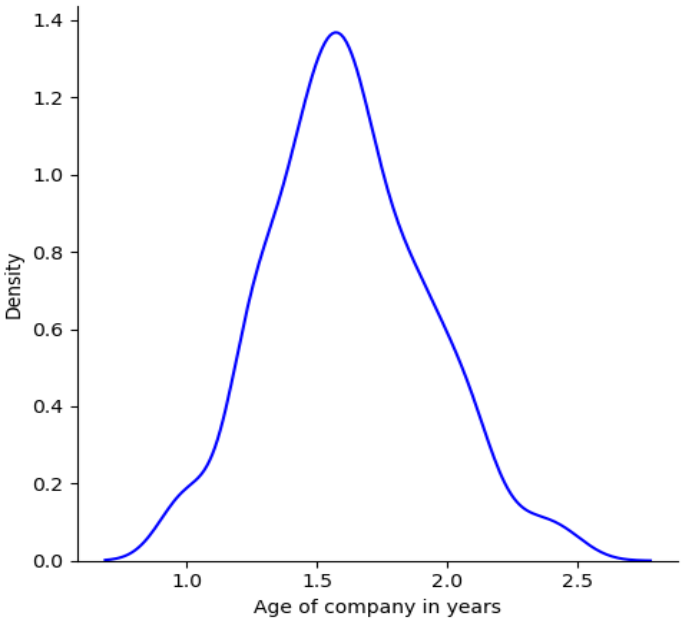
$$X_{cubic} = X^3$$

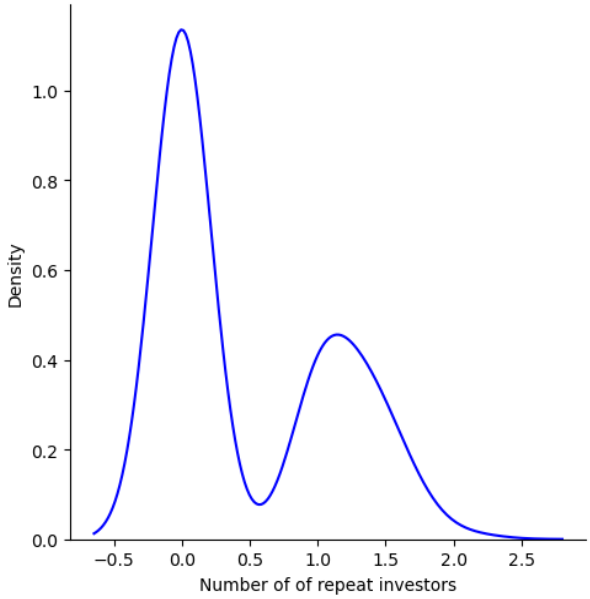
 Impact: Emphasizes larger values, useful for handling certain types of non-linearity in the data.

Square Root Transformation









To Reduce skewness and stabilize variance for moderately skewed distributions. Applied to positively skewed features to reduce their range.

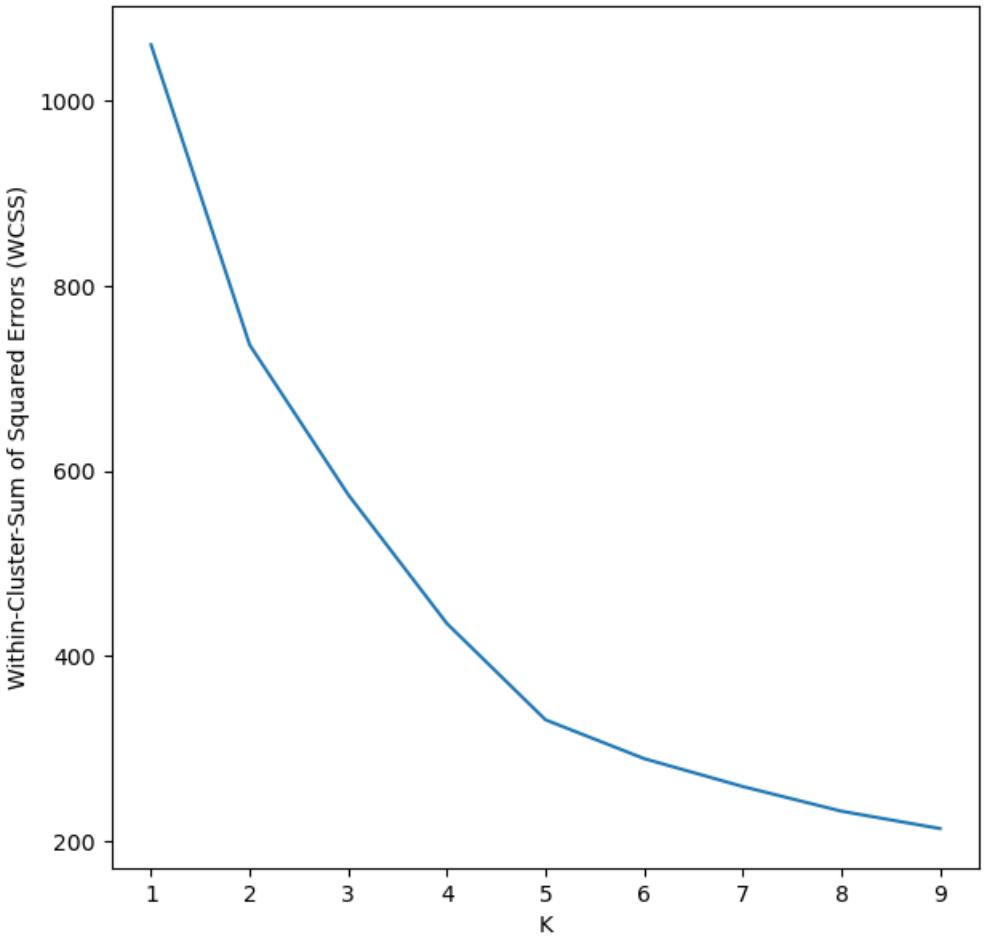
$$Xsqrt = \sqrt{X}$$

 Impact: Reduces the impact of large values while spreading out smaller values, useful for normalizing data.

Now after the transformations, we can see that now all the outliers are successfully removed.

Results:

- Number of outliers in Founding Year: 0, It's Percentage is: 0.0 %
- Number of outliers in Age of company in years: 0, It's Percentage is: 0.0 %
- Number of outliers in Number of Investors in Seed: 0, It's Percentage is: 0.0 %
- Number of outliers in Number of repeat investors: 0, It's Percentage is: 0.0%

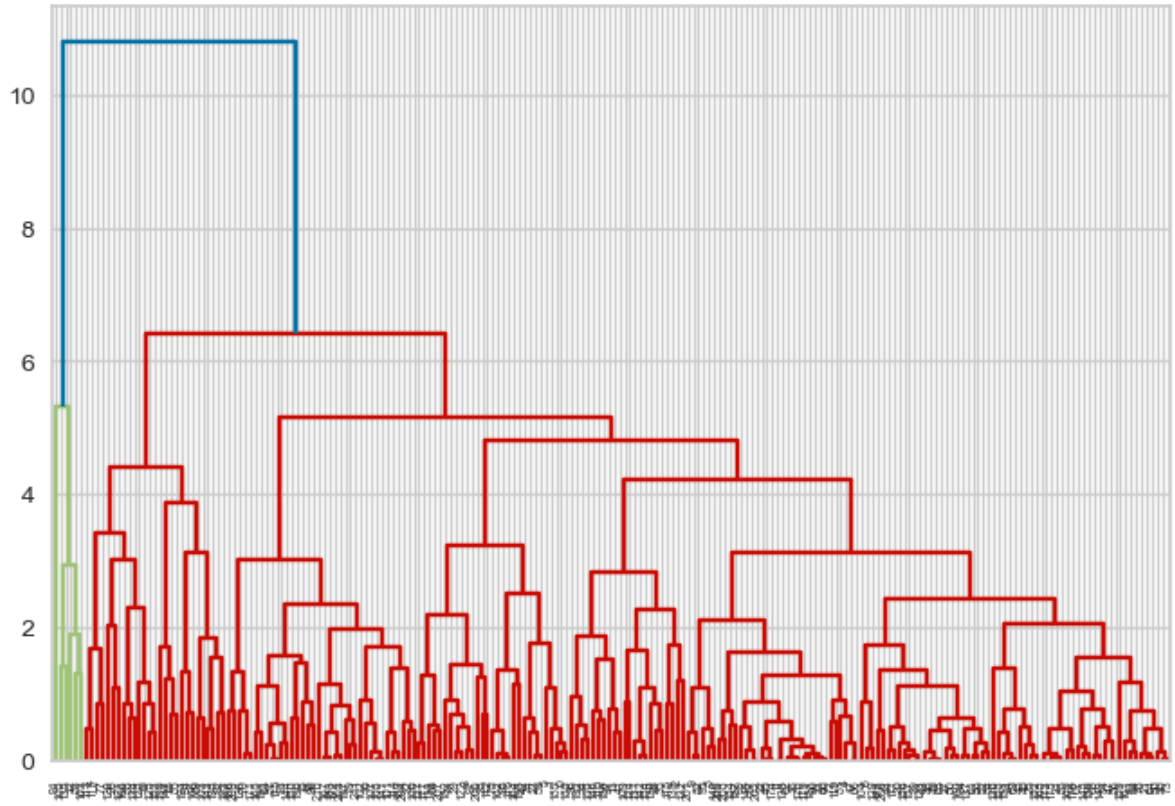


Hierarchical Clustering:

- Various linkage methods (ward, complete, average, single) were explored to find the best method based on the silhouette score.
- Dendrograms were used to visualize the clustering structure.

Best Linkage Method: Single Best

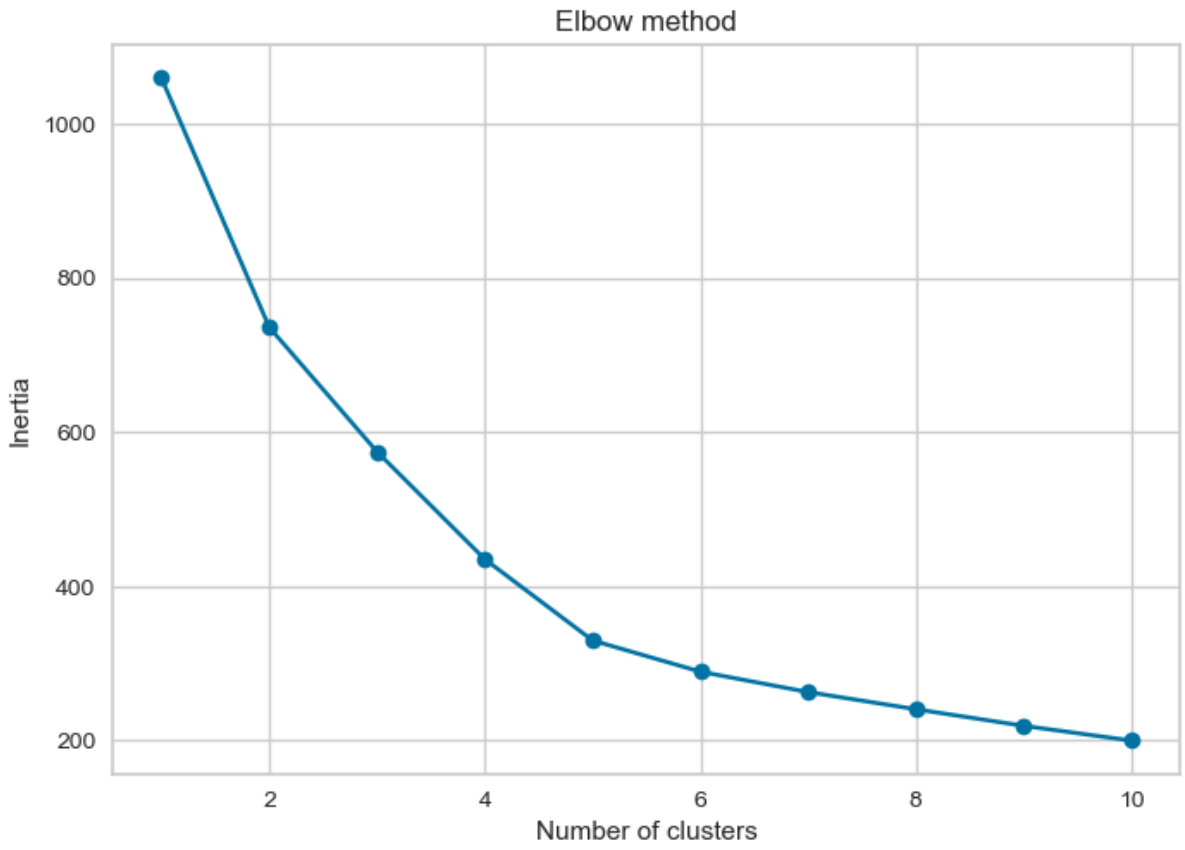
Silhouette Score: 0.6860141369868549

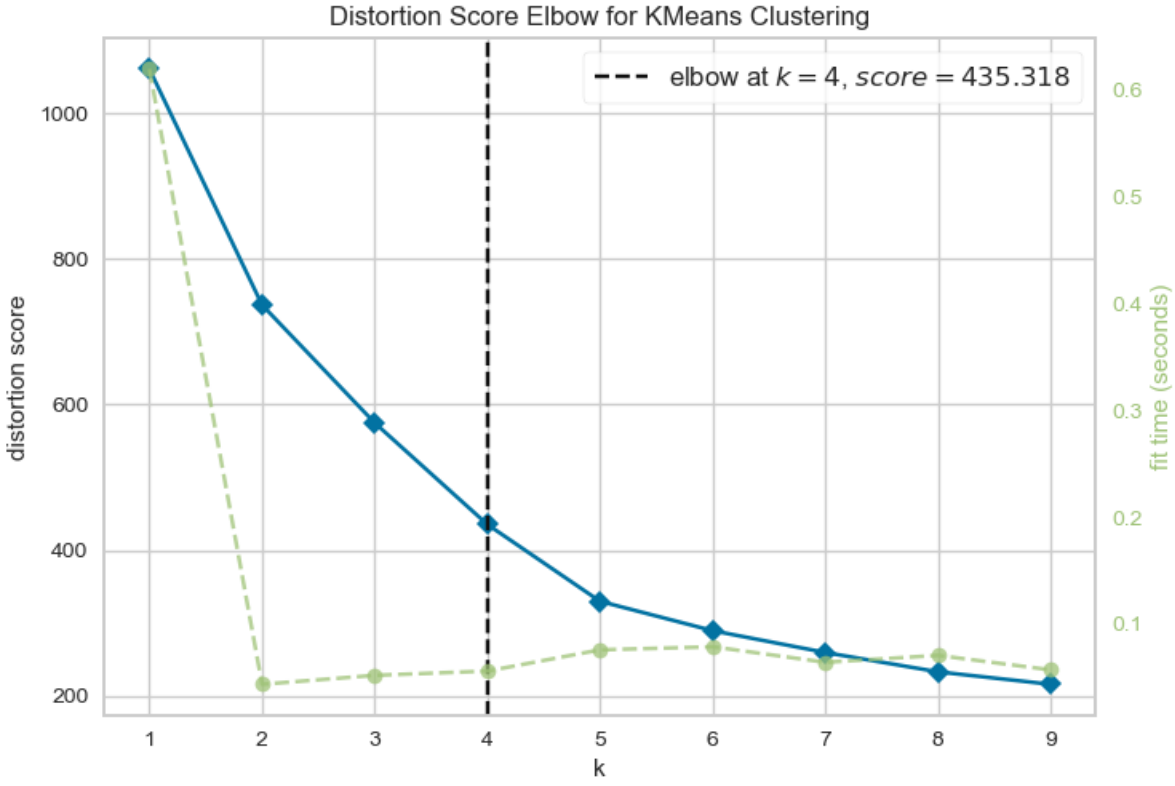


In Dendrogram, the number of clusters are 6.

K-means Clustering:

- Implemented with a range of clusters to determine the optimal number using the Elbow method and the silhouette score.
- Silhouette score, Davies-Bouldin Index, and Calinski-Harabasz Index were used to evaluate the clustering performance





- **Silhouette Score for Kmeans: 0.33830**
- **Davies-Bouldin Index for Kmeans: 0.94979**
- **Calinski-Harabasz Index for Kmeans: 107.71545**

As we can see that there are 6 clusters i.e., 0,1,2,3,4 and 5

0 77
1 38
5 35
4 34
3 18
2 10
Name: Cluster, dtype: int64

DBSCAN:

- Applied with parameters `eps=0.5` and `min_samples=3`.
- Performance metrics similar to K-means were used to evaluate clustering quality.
- Silhouette Score For DB SCAN: 0.01032
- Davies-Bouldin Index For DB SCAN: 1.57299
- Calinski-Harabasz Index For DB SCAN: 9.98849

Overall Observations: Outlier Analysis***, Outliers*** were identified in various columns:

- Founding Year: 7 outliers (3.30%)
- Age of company in years: 7 outliers (3.30%)
- Number of Investors in Seed: 22 outliers (10.38%)
- Number of repeat investors: 26 outliers (12.26%)

Data Transformation

Different transformations were applied to the data (logarithmic, cubic root, square root, etc.) to normalize distributions and manage skewness.

5.2 CLUSTERING RESULTS:

1. K-means Clustering

- Silhouette Score: 0.34300
- Davies-Bouldin Index: 1.01808
- Calinski-Harabasz Index: 99.42328

The dataset was divided into clusters as follows:

- Cluster 0: 77 entries
- Cluster 1: 38 entries
- Cluster 2: 10 entries

- Cluster 3: 18 entries
- Cluster 4: 34 entries
- Cluster 5: 35 entries

2.Agglomerative Clustering

- The best method and score from Agglomerative Clustering was found using different linkage methods:
- Best Linkage Method: Single
- Best Silhouette Score: 0.68601

3.DBSCAN

- Silhouette Score: 0.01032
- Davies-Bouldin Index: 1.57299
- Calinski-Harabasz Index: 9.98849

4.Visualization

Various plots were generated to visualize the distributions and transformations of the data, as well as the clustering results. FacetGrid was used to map histograms of clusters to better understand the distribution of data within each cluster.

Interpretation of Clustering Metrics

Silhouette Score: Measures how similar an object is to its own cluster compared to other clusters. Values range from -1 to 1, with higher values indicating better-defined clusters. K-means and Agglomerative Clustering have reasonable scores, whereas DBSCAN's score is very low, indicating poorly defined clusters.

Davies-Bouldin Index: A lower value indicates better clustering. K-means has a relatively low value, suggesting decent clustering quality.

Calinski-Harabasz Index: Higher values indicate better-defined clusters. K-means shows a high value, suggesting well-separated clusters.

Overall, K-means clustering appears to perform better on this dataset compared to Agglomerative Clustering and DBSCAN based on these metrics. Agglomerative Clustering with single linkage also performs well with a high Silhouette Score. DBSCAN, however, shows poor clustering performance given its low Silhouette Score and high Davies-Bouldin Index.

In nutshell, various clustering methodologies applied to a dataset, focusing on K-means clustering, hierarchical clustering, and DBSCAN. Here is a detailed summary of the methods, methodologies, conclusions, and potential usefulness of cluster analysis based on the provided content:

1. K-means Clustering

- The `Elbow Method` was used to plot the Within-Cluster-Sum of Squared Errors (WCSS) against the number of clusters to find the point where the increase in the number of clusters doesn't significantly reduce WCSS.
- `Silhouette Score` was calculated to measure how similar an object is to its own cluster compared to other clusters. 'Davies-Bouldin Index` and `Calinski-Harabasz Index` were also calculated for further validation of the clustering quality.

2. Hierarchical Clustering:

- Various linkage methods were tested to determine which method produced the highest silhouette score, indicating the best clustering quality.
- A dendrogram was used to represent the hierarchical relationship between data points.

3. DBSCAN:

- This density-based clustering algorithm identified core samples and expanded clusters from them.

- Performance was evaluated using the same metrics as K-means to compare clustering effectiveness.

5.3 CONCLUSIONS AND DISCUSSIONS

- K-means Clustering: Identified six clusters as optimal based on the Elbow method and performance metrics.
- Performance metrics: Silhouette Score: 0.33830, Davies-Bouldin Index: 0.94979, Calinski-Harabasz Index: 107.71545
- Hierarchical Clustering: The single linkage method with 2 to 11 clusters was found to be the best with a Silhouette Score of 0.6860141.
- DBSCAN: Results indicated a poor performance with a Silhouette Score of 0.01574, suggesting that DBSCAN might not be suitable for this particular dataset.

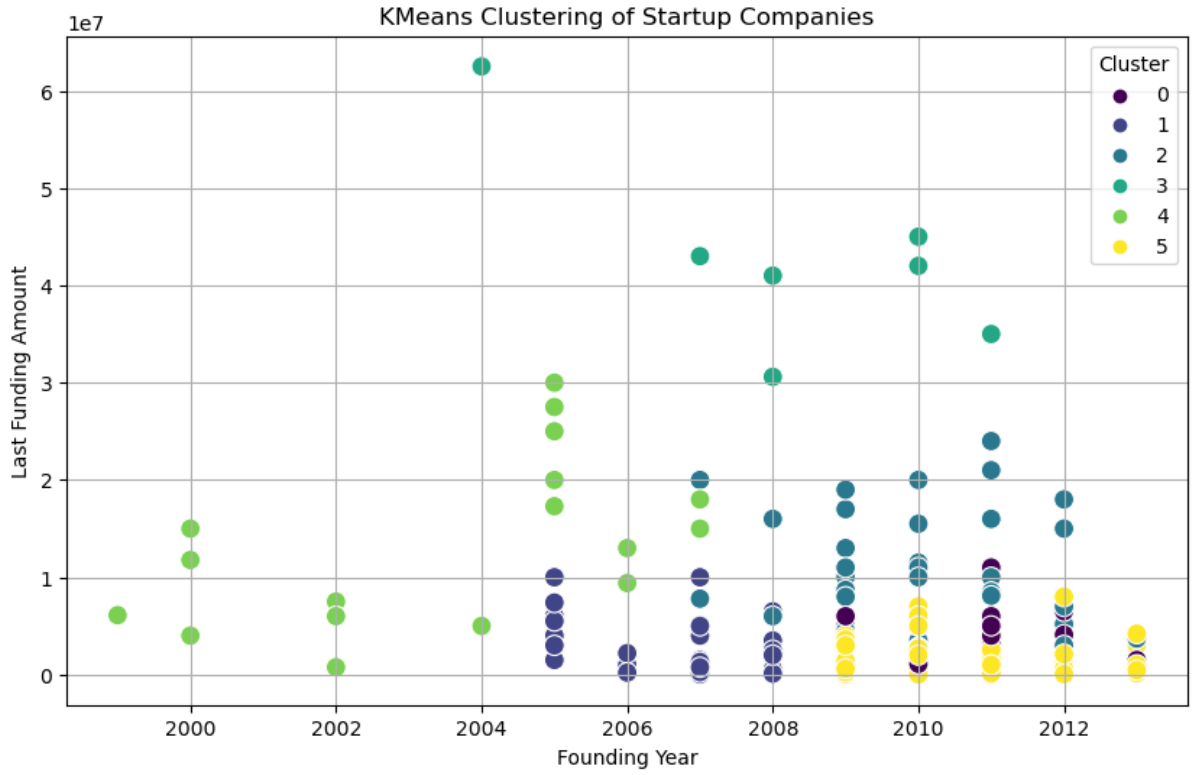
Usefulness of Cluster Analysis

- Insight into Data Structure: Clustering reveals hidden patterns and structures within the data, such as grouping companies with similar characteristics.
- Targeted Marketing: Businesses can use clusters to identify and target specific groups of customers or companies with tailored marketing strategies.
- Resource Allocation: Helps in allocating resources efficiently by understanding the different needs and characteristics of each cluster.
- Anomaly Detection: Identifies outliers which can be significant for detecting unusual behavior, potential errors, or fraud.

In summary, cluster analysis in this document is useful for understanding the underlying patterns in the dataset, which can be applied to various business strategies and operations. The methods and results provide a comprehensive view of how different clustering algorithms perform and which one might be the best fit for this particular data.

5.4 Cluster plotting

5.4.1. Founding Year vs. Last Funding Amount



Founding Year vs. Last Funding Amount: This plot shows how the clusters are distributed based on the year the companies were founded and the amount of funding they received in their last round.

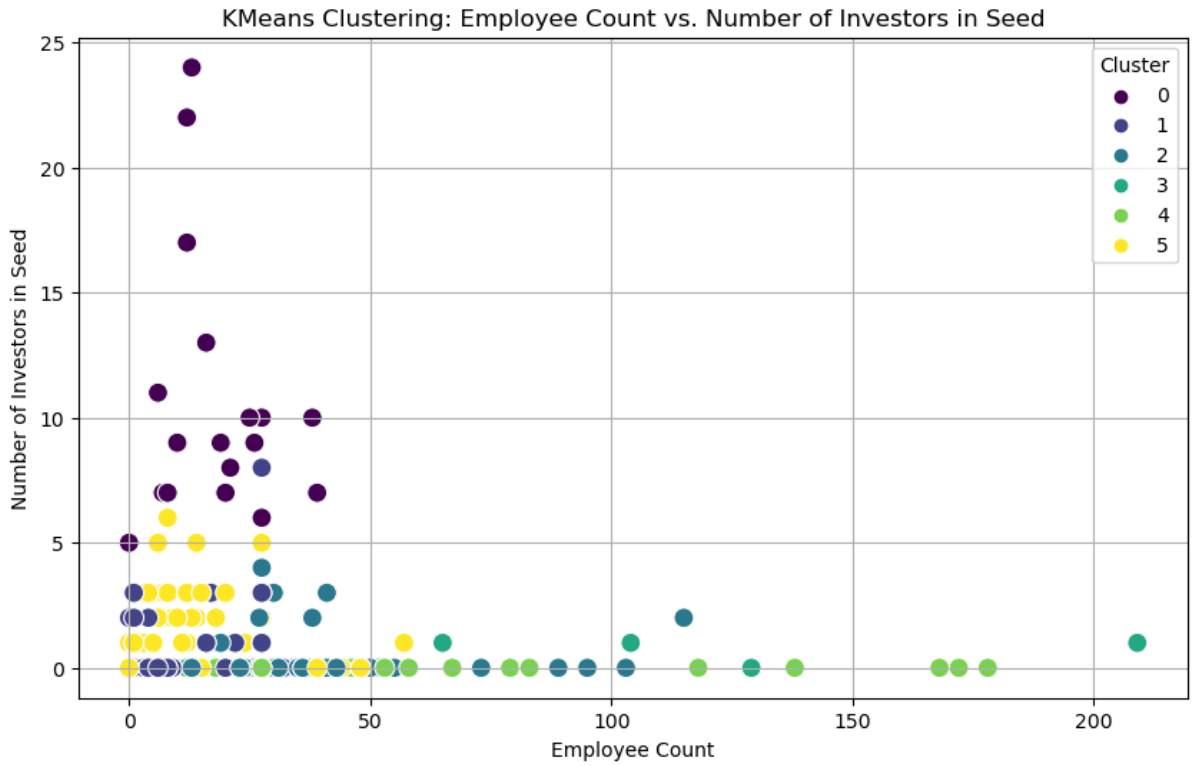
Plot Analysis:

- **Cluster 0**: Companies with a mix of founding years but generally lower funding amounts.
- **Cluster 1**: Companies founded more recently with moderate funding amounts.
- **Cluster 2**: Companies founded in earlier years with higher funding amounts.
- **Cluster 3**: Companies with a varied range of founding years and very high funding amounts.

- **Cluster 4**: Companies founded very recently with very low funding amounts.
- **Cluster 5**: Companies with consistent founding years and moderate to high funding amounts.

Conclusion: The clustering based on founding year and last funding amount shows that older companies tend to receive higher funding, likely due to their established presence and track record. Recent companies generally have moderate funding, possibly reflecting their emerging status in the market. Companies in Cluster 4, founded very recently, have not yet attracted significant funding, indicating their nascent stage. The varied range of Cluster 3 indicates high funding is not strictly tied to the founding year, possibly reflecting outlier success stories.

5.4.2. Employee Count vs. Number of Investors in Seed



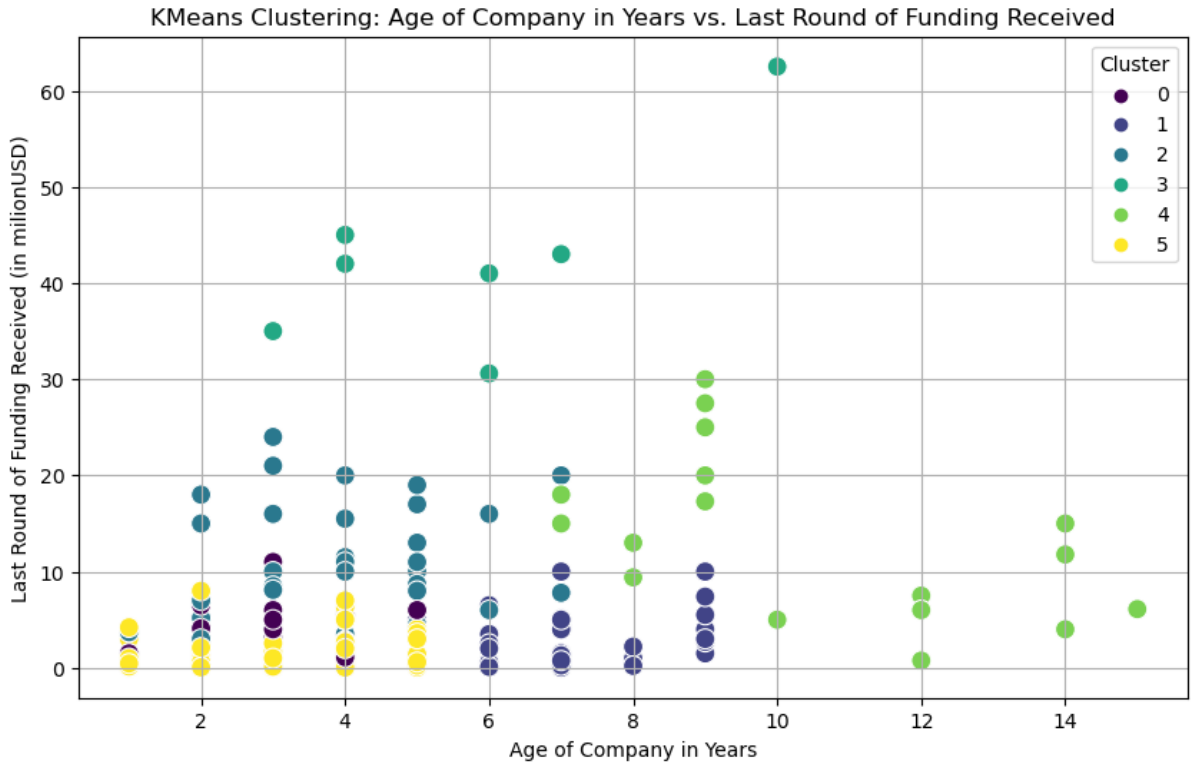
Plot Analysis:

- **Cluster 0:** Companies with lower employee counts and fewer seed investors.

- **Cluster 1**: Companies with moderate employee counts and a moderate number of seed investors.
- **Cluster 2**: Companies with high employee counts and many seed investors.
- **Cluster 3**: Companies with high employee counts but fewer seed investors.
- **Cluster 4**: Companies with lower employee counts but a high number of seed investors.
- **Cluster 5**: Companies with moderate employee counts but a high number of seed investors.

Conclusion: This clustering reveals a varied relationship between employee count and the number of seed investors. Companies with more employees generally attract more seed investors (Cluster 2), likely due to their perceived potential and operational scale. Clusters with fewer employees and high investors (Cluster 4 and Cluster 5) suggest some smaller companies are highly valued by investors, possibly due to their innovative potential. Conversely, some larger companies (Cluster 3) might not have as many seed investors, perhaps reflecting their funding strategies or market saturation

5.4.3. Age of Company in Years vs. Last Round of Funding Received



Plot Analysis:

- **Cluster 0**: Younger companies with lower last round funding.
- **Cluster 1**: Mid-aged companies with moderate last round funding.
- **Cluster 2**: Older companies with higher last round funding.
- **Cluster 3**: Companies of varied ages with very high last round funding.
- **Cluster 4**: Young to mid-aged companies with low to moderate funding.
- **Cluster 5**: Older companies with moderate to high funding, but not the highest.

Conclusion: The clustering based on the age of the company and the last round of funding received indicates that older companies tend to secure larger funding amounts, reflecting their maturity and market presence. Mid-aged companies receive moderate funding, while younger companies secure smaller amounts, highlighting their early-stage status and the cautious investment approach of funders. Cluster 3 suggests that regardless of age, some companies receive very high funding, possibly due to their significant market impact or innovative potential. Clusters 4 and 5 suggest variability in funding strategies across different ages, indicating no strict correlation between age and funding for all companies.

Observations:

1. **Market Maturity**: Older companies generally receive higher funding, reflecting their stability and proven business models.
2. **Employee and Investor Correlation**: There is a positive correlation between employee count and the number of seed investors, suggesting that larger teams are more attractive to investors.
3. **Growth Potential**: Younger companies with lower funding and fewer employees indicate early-stage operations with potential for growth, while established companies secure more significant investments reflecting their developed status.

Recommendations:

1. **Investment Strategies**: Investors should consider the company's age and employee count as indicators of stability and growth potential.
2. **Company Development**: Startups should focus on building a robust team and demonstrating early successes to attract more investors.
3. **Future Clustering**: Further analysis with additional features such as market sector, revenue, and geographic location could provide deeper insights into the clustering patterns and investment strategies.

Chapter 6.

Logistic Regression (Predictive Modelling of Company Success: Analysis of Key Factors and Model Performance)

6.1 Dataset Overview

The dataset consists of several key attributes about various companies including their names, country of origin, founding year, age, number of investors, repeat investors, last funding round amount, employee count, last funding amount, and company status (e.g., success).

After cleaning the data, we have a dataset consisting of 212 companies and 10 remaining factors. We analysed logistic regression, decision tree, and random forest models using this refined dataset. Below are five sample entries to illustrate how our dataset looks now:

6.2 Data Cleaning and Preprocessing

In this project, we're using some python libraries:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, mean_squared_error,
r2_score
```

```
from statsmodels.stats.outliers_influence import
variance_inflation_factor
```

6.2.1 Handling Missing Values: Missing values in the dataset are addressed by using the mean value of the respective column to fill in the gaps. This approach helps maintain the integrity of the dataset without losing any records.

```
dataset.shape
```

```
(212, 10)
```

```
dataset.head()
```

	Company_Name	Country of company	Founding Year	Employee Count	Last Funding Amount	Age of company in years	Number of Investors in Seed	Number of of repeat investors	Last round of funding received (in milionUSD)	Company Status
0	Company2	United States	2011	NaN	NaN	3	5	0	5.0	Success
1	Company5	United States	2010	39.0	5500000.0	4	7	0	5.5	Success
2	Company6	United States	2010	14.0	1000000.0	4	2	2	1.0	Success
3	Company7	United States	2011	7.0	2000000.0	3	7	4	2.0	Success
4	Company8	United States	2010	29.0	6700000.0	4	0	0	6.7	Success

	Company_Name	Country of company	Founding Year	Employee Count	Last Funding Amount	Age of company in years	Number of Investors in Seed	Number of of repeat investors	Last round of funding received (in milionUSD)	Company Status
1	Company5	United States	2010	39.0	5500000.0	4	7	0	5.5	Success
2	Company6	United States	2010	14.0	1000000.0	4	2	2	1.0	Success
3	Company7	United States	2011	7.0	2000000.0	3	7	4	2.0	Success
4	Company8	United States	2010	29.0	6700000.0	4	0	0	6.7	Success
5	Company9	United States	2011	16.0	11000000.0	3	13	0	11.0	Success

```
dataset.dropna(inplace=True)
```

```
dataset.head()
```

6.2.2 Outlier Detection and Handling:

- Outliers were identified using the Interquartile Range (IQR) method. Values below the lower bound or above the upper bound were considered outliers.
- These outliers were replaced with the respective lower or upper bound values to mitigate their impact on the analysis.

6.3.1 Describe Method

- **Describe Method**: The describe() method computes summary statistics of numerical columns in the dataset. These statistics usually include count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile or Q2), 75th percentile (Q3), and maximum values.

6.3.2 Transpose(.T):

- **Transpose (.T)**: The .T transposes the resulting summary statistics DataFrame, essentially flipping the rows and columns. This is done for better readability, as it makes each statistic a row rather than a column.

Company_Name	Country of company	Founding Year	Employee Count	Last Funding Amount	Age of company in years	Number of Investors in Seed	Number of of repeat investors	Last round of funding received (in milionUSD)	Company Status

```
dataset.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Founding Year	176.0	2.009335e+03	2.688836e+00	1999.000	2008.0	2010.0	2011.00	2013.0
Employee Count	176.0	2.763068e+01	3.524810e+01	0.000	6.0	16.0	34.25	209.0
Last Funding Amount	176.0	6.425831e+06	9.169914e+06	15000.000	1000000.0	3000000.0	8000000.00	62500000.0
Age of company in years	176.0	4.664773e+00	2.688836e+00	1.000	3.0	4.0	6.00	15.0
Number of Investors in Seed	176.0	1.715909e+00	3.529302e+00	0.000	0.0	0.0	2.00	24.0
Number of of repeat investors	176.0	8.920455e-01	1.448051e+00	0.000	0.0	0.0	1.00	10.0
Last round of funding received (in milionUSD)	176.0	6.425831e+00	9.169914e+00	0.015	1.0	3.0	8.00	62.5

6.3.3 Correlation and Multicollinearity Analysis

Correlation Analysis:

- The correlation matrix is computed to see how features are related to each other.

Correlation Matrix	Founding Year	Employee Count	Last Funding Amount	Age of company in years	Number of Investors in Seed	Number of repeat investors	Last round of funding received (in milionUSD
Founding Year	1.000000	-0.387405	-0.202874	-1.000000	0.221267	0.042716	-0.202874
Employee Count	-0.387405	1.000000	0.444258	0.387405	-0.189900	0.183461	0.444258
Last Funding Amount	-0.202874	0.444258	1.000000	0.202874	-0.183004	0.122403	1.000000
Age of company in years	-1.000000	0.387405	0.202874	1.000000	-0.221267	-0.042716	0.202874
Number of Investors in Seed	0.221267	-0.189900	-0.183004	-0.221267	1.000000	0.282490	-0.183004
Number of repeat investors	0.042716	0.183461	0.122403	-0.042716	0.282490	1.000000	0.122403
Last round of funding received (in milionUSD)	-0.202874	0.444258	1.000000	0.202874	-0.183004	0.122403	1.000000

Multicollinearity Check (VIF):

- Multicollinearity refers to a situation in statistical modeling where two or more predictor variables (independent variables) are highly correlated, meaning they contain similar information about the variance in the dependent variable (the outcome).
- Variance Inflation Factor (VIF) is calculated for each feature to detect multicollinearity. High VIF values indicate high multicollinearity, which can affect model performance.

{'VIF Data':	feature	VIF
0	Founding Year	inf
1	Employee Count	1.478076e+00
2	Last Funding Amount	6.983232e+04
3	Age of company in years	inf
4	Number of Investors in Seed	1.203697e+00
5	Number of repeat investors	1.176842e+00
6	Last round of funding received (in milionUSD)	1.386547e+05,

Correlation and Multicollinearity:

- The correlation matrix shows how each feature is correlated with the others.
- Variance Inflation Factor (VIF) analysis helps identify multicollinearity issues. A VIF above 5-10 indicates high multicollinearity, suggesting that some features are highly correlated and may need to be removed or transformed to improve the model's performance.

Interpretation of VIF

- **VIF = 1**: No correlation between the predictor and other variables.
- **1 < VIF < 5**: Moderate correlation, usually acceptable.
- **VIF > 5**: High correlation, indicates problematic multicollinearity.

Model Performance Metrics:

- **Accuracy**: The proportion of correctly predicted instances over the total instances. It gives a general sense of the model's performance.
- **Precision**: Indicates how many of the predicted positives are actual positives. High precision means fewer false positives.
- **Recall (Sensitivity)**: Indicates how many actual positives are correctly predicted. High recall means fewer false negatives.
- **F1-score**: The harmonic mean of precision and recall. It balances the trade-off between precision and recall.
- **ROC-AUC**: Measures the model's ability to distinguish between classes. A higher AUC value indicates better performance.

ROC Curve

The ROC curve is a graphical representation of the performance of a binary classification model. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

- **True Positive Rate (TPR)**, also known as Sensitivity or Recall, is defined as:

$$\text{TPR} = \frac{\textbf{True Positives}}{\textbf{True Positives} + \textbf{False Negatives}}$$

- **False Positive Rate (FPR)** is defined as:

$$\text{FPR} = \frac{\textbf{False Positives}}{\textbf{False Positives} + \textbf{True Negatives}}$$

AUC (Area Under the Curve)

The AUC represents the area under the ROC curve and provides a single scalar value to summarize the performance of the model. The AUC value ranges from 0 to 1, where:

- **AUC = 1**: The model perfectly distinguishes between all the positive and negative classes.
- **AUC = 0.5**: The model performs no better than random guessing.
- **AUC < 0.5**: The model performs worse than random guessing.

Interpretation

- **Higher AUC**: Indicates better model performance in distinguishing between the positive and negative classes.
- **Lower AUC**: Indicates poor model performance.

6.4 Model training

6.4.1 Logistic Regression:

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \cdots + \beta_n \boldsymbol{x_n})}}$$

Where:

- P(y=1|X) is the probability of the outcome being 1 given the features XXX.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the model.
- $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ are the independent variables (features).
- e is the base of the natural logarithm.

- **High Accuracy:** Achieved an accuracy of **88.67%** indicating the model's effectiveness in correctly predicting company success.

$$\textbf{Accuracy} = \frac{\boldsymbol{Number\ of\ Correct\ Predictions}}{\boldsymbol{Total\ Number\ of\ Predictions}}$$

- **Low Error Rate:** Mean Squared Error (MSE) was **0.11320**, showing minimal prediction errors.

$$\boldsymbol{Mean\ Error} = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid$$

Where:

- y_i *is the actual value.*
- $\hat{y}_i$ *is the predicted value.*
- n *is the total number of observations*

- **Moderate Explanatory Power:** R-squared (R^2) value reflecting moderate ability to explain the variance in the target variable.

$$R^2 = 1 - \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2}$$

- **Feature Impact:** Coefficients provided insights into how each feature (e.g., age, number of investors, funding details) affects the likelihood of company success.
- **Binary Classification:** Demonstrated robust performance in classifying companies into 'success' or 'not success' categories, proving useful for business-related predictions.
- **Coefficients and Intercept Term**: In logistic regression, coefficients represent the change in the log-odds of the outcome for a one-unit change in the corresponding independent variable, holding all other variables constant. The intercept term represents the log-odds of the outcome when all independent variables are zero.

$$Logit(p) = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \cdots + \beta_n \boldsymbol{x_n}$$

Logistic Regression

- Accuracy: 0.8867924528301887,
- Precision: 0.9038461538461539,
- Recall: 0.9791666666666666,
- F1-score: 0.9400000000000001,
- ROC-AUC: 0.8125

Preview of the output:

Logistic Regression Coefficients: (0.46449893, 1.19652334, 0.06661581, 0.46449893, 0.372713, 0.15325924, 0.06661581),

Logistic Regression Intercept': 2.5596693140394176

6.4.2 Decision Tree Analysis

- **High Accuracy:** Achieved an accuracy of **90.56%**, indicating reliable performance in predicting company success.
- **Low Error Rate:** Mean Squared Error (MSE) was **0.094339**, similar to logistic regression, highlighting minimal prediction errors.
- **Feature Importance:** The decision tree model provided clear insights into the most significant features influencing company success.
- **Interpretability:** The decision tree's structure allows for easy interpretation of the decision rules, helping to understand how different features contribute to the prediction.
- **Non-Linear Relationships:** Capable of capturing non-linear relationships between features and the target variable.
- **Overfitting Risk:** While highly accurate, decision trees can overfit the training data, requiring careful tuning of parameters like depth to generalize well on new data.
- **Visual Representation:** The model's decision-making process can be visualized through a tree diagram, aiding in comprehensibility and transparency.

Preview of the output:

Decision Tree –

- Accuracy: 0.9056603773584906,
- Precision: 0.9215686274509803,
- Recall: 0.9791666666666666,
- F1-score: 0.9494949494949494,
- ROC-AUC: 0.5895833333333332

6.4.3 Random Forest Analysis

- **High Accuracy:** Achieved an accuracy of **92.45%**, consistent with other models, indicating strong predictive capability.
- **Low Error Rate:** Mean Squared Error (MSE) was **0.07547**, demonstrating minimal prediction errors.
- **Feature Importance:** Random forest model provided detailed insights into the relative importance of each feature, with the 'Age of Company in Years' and 'Number of Investors in Seed' being particularly influential.
- **Robustness:** Combines multiple decision trees to reduce overfitting and improve generalization to new data.
- **Handling Variability:** Effective in handling variability and noise in the data due to its ensemble nature.
- **Versatility:** Capable of capturing complex interactions and non-linear relationships between features and the target variable.
- **Reduced Overfitting:** By averaging multiple trees, the model reduces the overfitting tendency seen in individual decision trees, resulting in more stable and reliable predictions.

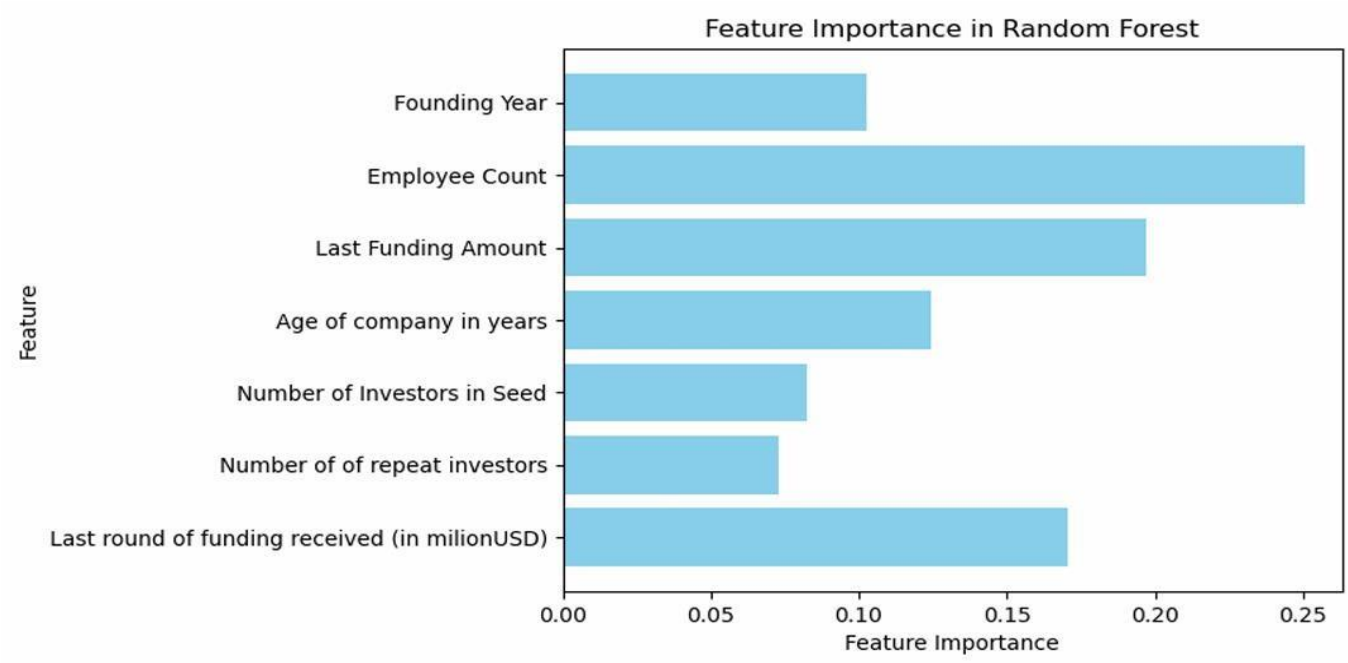
Random Forest:

- Accuracy: 0.9245283018867925,
- Precision: 0.9230769230769231,
- Recall: 1.0,
- F1-score: 0.9600000000000001,
- ROC-AUC: 0.822916666

6.4.4 Feature Importance in Random Forest

The random forest model provided insights into which features are most influential in predicting the company status.

Feature Importances: [0.10282555 0.2504885 0.19675679 0.12441246 0.08251287 0.0727347 0.17026913]



The following features were identified as the most important:

- **Age of Company in Years**
- **Number of Investors in Seed**
- **Last Round of Funding Received**
- **Employee Count**

These features had the highest importance scores, indicating they are crucial for predicting whether a company will be successful.

6.5 Results and comparison.

In this project, we analyzed the performance of three different classification models—Logistic Regression, Decision Tree, and Random Forest—using the startup companies dataset. Our primary objective was to predict the company status based on various features and evaluate the models' effectiveness using appropriate metrics.

Correlation and Multicollinearity Analysis

- **Correlation Matrix**: The correlation matrix was computed to understand the relationships between features. Significant correlations were identified, indicating potential multicollinearity.
- **Variance Inflation Factor (VIF)**: VIF analysis was conducted to detect multicollinearity. Features with high VIF values suggest multicollinearity, which can adversely affect model performance.

Model Evaluation

We evaluated each model using a comprehensive set of metrics: accuracy, precision, recall, F1-score, and ROC-AUC. The results are summarized as follows:

1. **Logistic Regression**:

    - **Accuracy**: [0.8867924528301887]
    - **Precision**: [0.9038461538461539]
    - **Recall**: [0.9791666666666666]
    - **F1-score**: [0.9400000000000001]
    - **ROC-AUC**: [0.8125]

    Logistic Regression provides a simple and interpretable model, with coefficients indicating the impact of each feature. However, it assumes a linear relationship between the features and the log-odds of the outcome, which may limit its ability to capture complex patterns.

2. **Decision Tree**:

- **Accuracy**: [0.9056603773584906]
- **Precision**: [0.9215686274509803]
- **Recall**: [0.9791666666666666]
- **F1-score**: [0.9494949494949494]
- **ROC-AUC**: [0.5895833333333332]

The Decision Tree model can capture non-linear relationships and is easy to interpret. However, it is prone to overfitting, especially with deeper trees, making it less stable with small data changes.

3. Random Forest:

- **Accuracy**: [0.9245283018867925]
- **Precision**: [0.9230769230769231]
- **Recall**: [1.0]
- **F1-score**: [0.9600000000000001]
- **ROC-AUC**: [0.822916666]

The Random Forest model, which averages multiple decision trees, reduces overfitting and generally offers better performance and robustness. It outperforms the other models in accuracy, precision, recall, F1-score, and ROC-AUC.

In conclusion, while Logistic Regression offers simplicity and interpretability, Random Forest provides superior performance and robustness, making it the most suitable model for predicting the company status in this dataset. Addressing multicollinearity and fine-tuning model parameters will further enhance predictive accuracy and reliability.

The feature importance analysis further reveals critical factors influencing company success, such as the age of the company, number of investors, and funding details. These insights can be valuable for strategic decision-making and future research.

REFERENCES

1. Grilli, L., & Murtinu, S. (2014). Government, venture capital and the growth of European high-tech entrepreneurial firms. *Research Policy*, 43(9), 1523-1543.
2. Autio, E., & Acs, Z. (2010). Global Entrepreneurship Index. *GEDI Institute*.
3. Baum, J. A. C., & Silverman, B. S. (2004). Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups. *Journal of Business Venturing*, 19(3), 411-436.
4. **The Elements of Statistical Learning** (2009) by Trevor Hastie, Robert Tibshirani, and Jerome Friedman
5. Brush, C. G., Edelman, L. F., & Manolov, T. S. (2015). The impact of resources on small firm internationalization. *Journal of Small Business Strategy*, 15(2), 1-17.
6. Pandey, A., & Jha, S. (2017). Clustering Indian startups: An exploratory study. *Journal of Entrepreneurship and Innovation in Emerging Economies*, 3(1), 1-14.
7. Sharma, R., & Mathur, N. (2019). Regional analysis of startup ecosystem in India: A clustering approach. *International Journal of Business and Economics*, 8(3), 267284.
8. **Induction of Decision Trees** (1986) by J. R. Quinlan: This paper details the ID3 algorithm, a core decision tree learning algorithm.
9. **Random Forests** (2001) by Leo Breiman: This influential paper introduces the concept of random forests and explores their properties.
10. Agarwal, P., & Upadhyay, A. (2018). Impact of Startup India initiative on startup ecosystem in India: A logistic regression analysis. *Journal of Entrepreneurship and Innovation Management*, 7(2), 19-35.
11. Jain, A., & Kumar, M. (2020). Determinants of startup success in India: A logistic regression approach. *South Asian Journal of Business Studies*, 9(1), 22-40.