Company-Clustering-using-Clustering_model(ML)

ASWIN.G.KUMAR

Introduction

In today's competitive business landscape, understanding company performance and customer satisfaction is crucial for driving strategic decision-making and identifying industry trends. The ability to cluster companies based on their ratings and characteristics can reveal valuable insights into their operational efficiencies, market positions, and customer engagement strategies.

This project aims to analyze a dataset containing information on 9,000 companies, focusing on their ratings and various attributes. By employing clustering techniques, we seek to group companies with similar characteristics, which will help in uncovering patterns that can inform business strategies and improve customer satisfaction.

Objectives:

1. **Clustering Companies:** Utilize unsupervised machine learning techniques to cluster companies based on their performance metrics, customer ratings, and other relevant characteristics.
2. **Pattern Identification:** Analyze the clusters to identify common traits among high-performing companies versus those with lower ratings. This analysis will help stakeholders understand the factors contributing to company success and customer satisfaction.
3. **Trend Analysis:** Explore industry trends based on the clusters, which can guide companies in benchmarking their performance against industry standards.

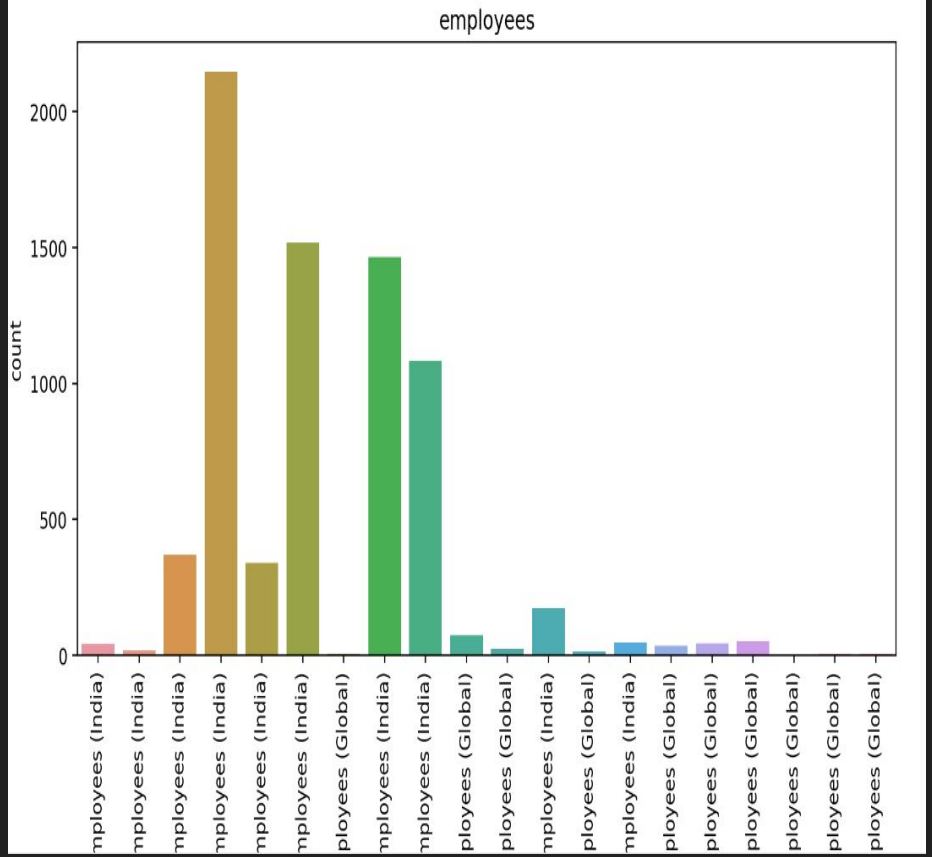
Tools and Technologies:

- **Scikit-learn:** A powerful machine learning library in Python that provides efficient tools for clustering, such as K-means and hierarchical clustering.
- **Pandas:** A data manipulation and analysis library that facilitates data cleaning, transformation, and exploration.
- **NumPy:** A library for numerical computations that will assist in handling large datasets and performing mathematical operations.
- **Matplotlib:** A visualization library to create insightful charts and graphs that illustrate the results of the clustering analysis.

Dataset:

The dataset is sourced from Kaggle and consists of comprehensive information about 9,000 companies, including their ratings and various attributes. This rich dataset serves as the foundation for our analysis, enabling us to apply clustering techniques effectively and derive meaningful insights

How to visualizes the distribution of companies based on the number of employees they have.

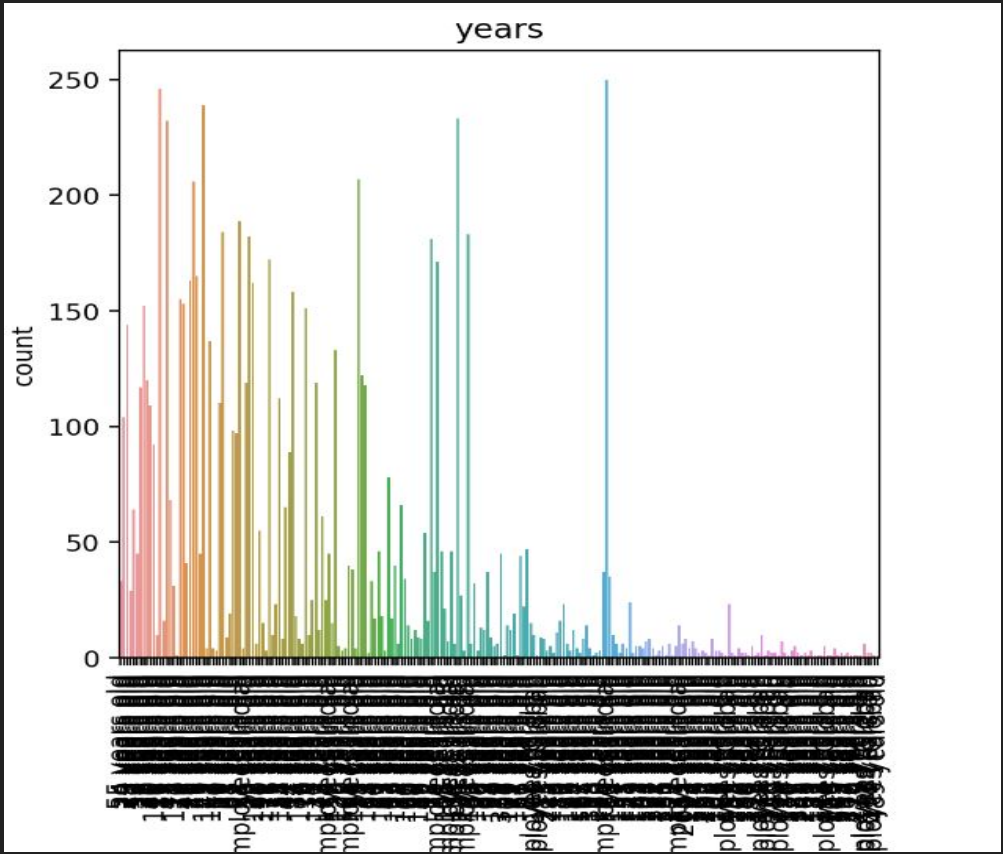


The bar chart you shared visualizes the distribution of companies based on the number of employees they have

- **X-axis (employees): The categories along the X-axis represent different employee ranges or types of employment (e.g., 1-10 employees, 11-50 employees, etc.). However, the labels are cluttered and overlapping, making it hard to read them clearly.**
- **Y-axis (count): The Y-axis shows the count of companies within each employee category. The height of each bar corresponds to how many companies fall within that range.**
- **Bars: Each bar represents the number of companies for a given employee range. From the chart:**

- **The largest number of companies appears to have around 100-200 employees.**
- **There are also significant numbers of companies with employees in other ranges, such as 51-200 or 201-500.**

How to visualize the distribution of companies based on the number of years they've been active



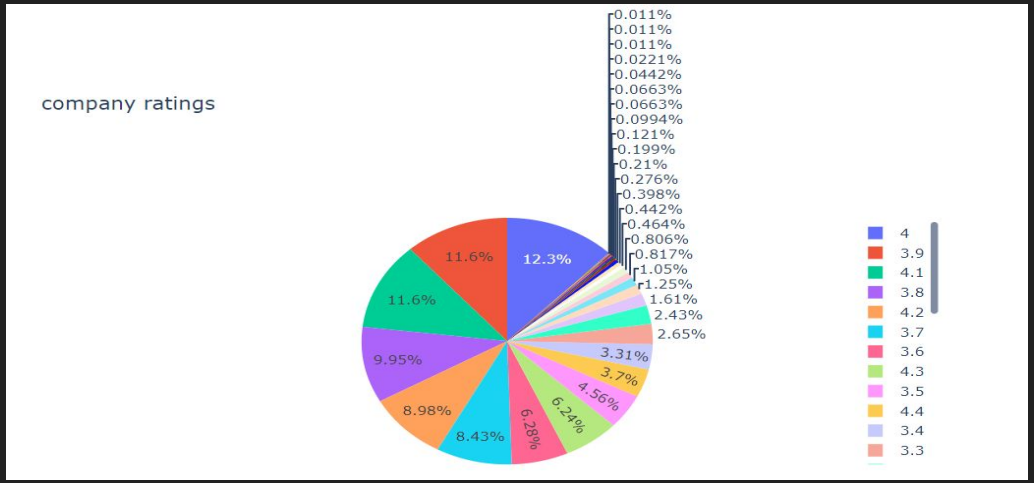
bar chart to visualize the distribution of companies based on the number of years they've been active or operating.

The X-axis will represent the number of years a company has been active (from the `years` column in your dataset).

The Y-axis will show the number of companies that correspond to each category of years.

The plot helps identify which years of company existence (e.g., 1-5 years, 6-10 years, etc.) are most common in the dataset.

How to find out the ratings of the companies

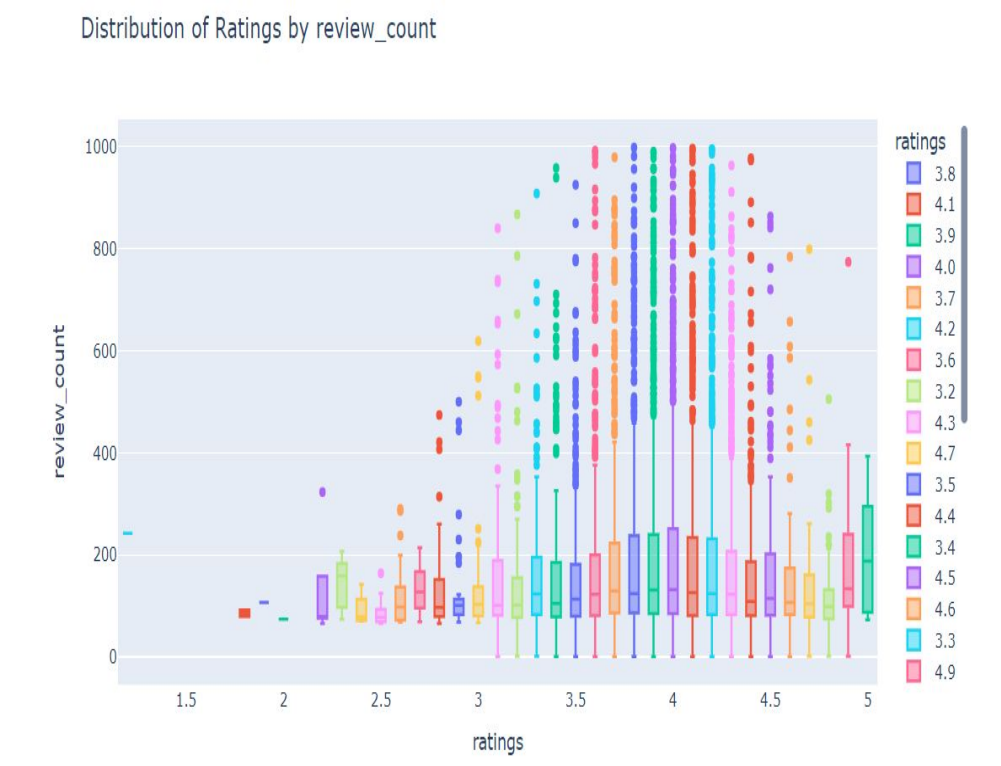


In this hist plot and pie chart shows the companies ratings step by step

The ratings are:

4, 3.9, 4.1, 3.8, 4.2, 3.7, 3.6, 4.3, 3.5, 4.4, 3.4 etc……

How to find the distribution of ratings by review_count



Here we create a box plot to find the distribution of ratings by review_counts

- Here the x value is ratings
- Here the Y value is review_count
-

Conclusion

Through clustering analysis of the 9,000 companies, we have successfully identified distinct groups of companies that share common characteristics and performance metrics. The patterns observed in customer satisfaction, industry trends, and overall ratings provide actionable insights for companies aiming to enhance their market positioning and operational efficiency. This clustering approach offers a foundation for deeper exploration into the drivers of success and failure across industries. Moreover, it equips businesses with the tools to identify potential benchmarks, areas for improvement, and opportunities for strategic growth within their respective clusters.

The dataset is sourced from Kaggle and consists of comprehensive information about 9,000 companies, including their ratings and various attributes. This rich dataset serves as the foundation for our analysis, enabling us to apply clustering techniques effectively and derive meaningful insights