CLUSTERING ANALYSIS ON GOOGLE PLAY STORE DATASET

Objective

This project focuses on performing a clustering analysis on the Google Play Store dataset to identify patterns and group similar applications together. The goal is to determine which clustering technique best fits the data set by comparing different algorithms and then analyze play store data using the particular technique.

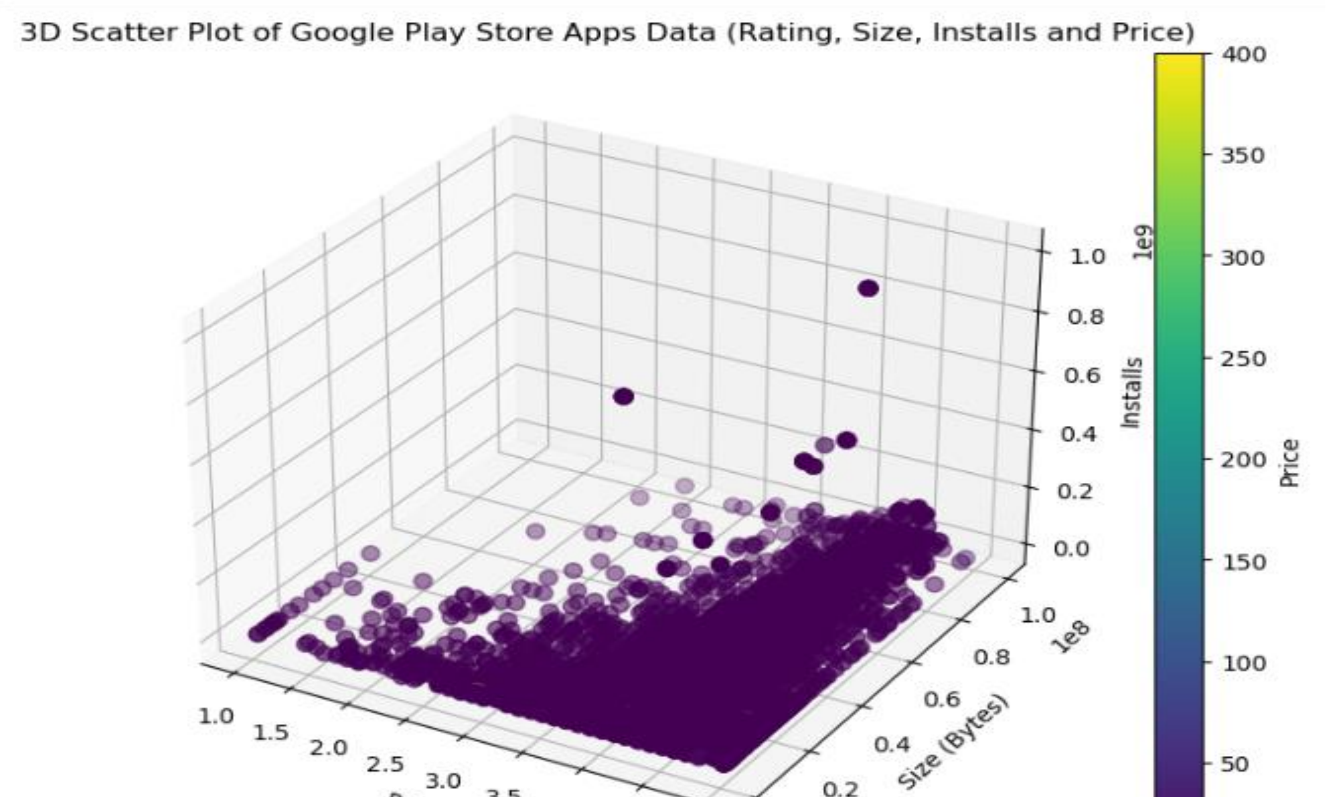
Introduction

Clustering is a type of unsupervised learning in machine learning where the goal is to group data points into distinct groups, or clusters, based on similarities in their features, to understand the underlying pattern and structure.

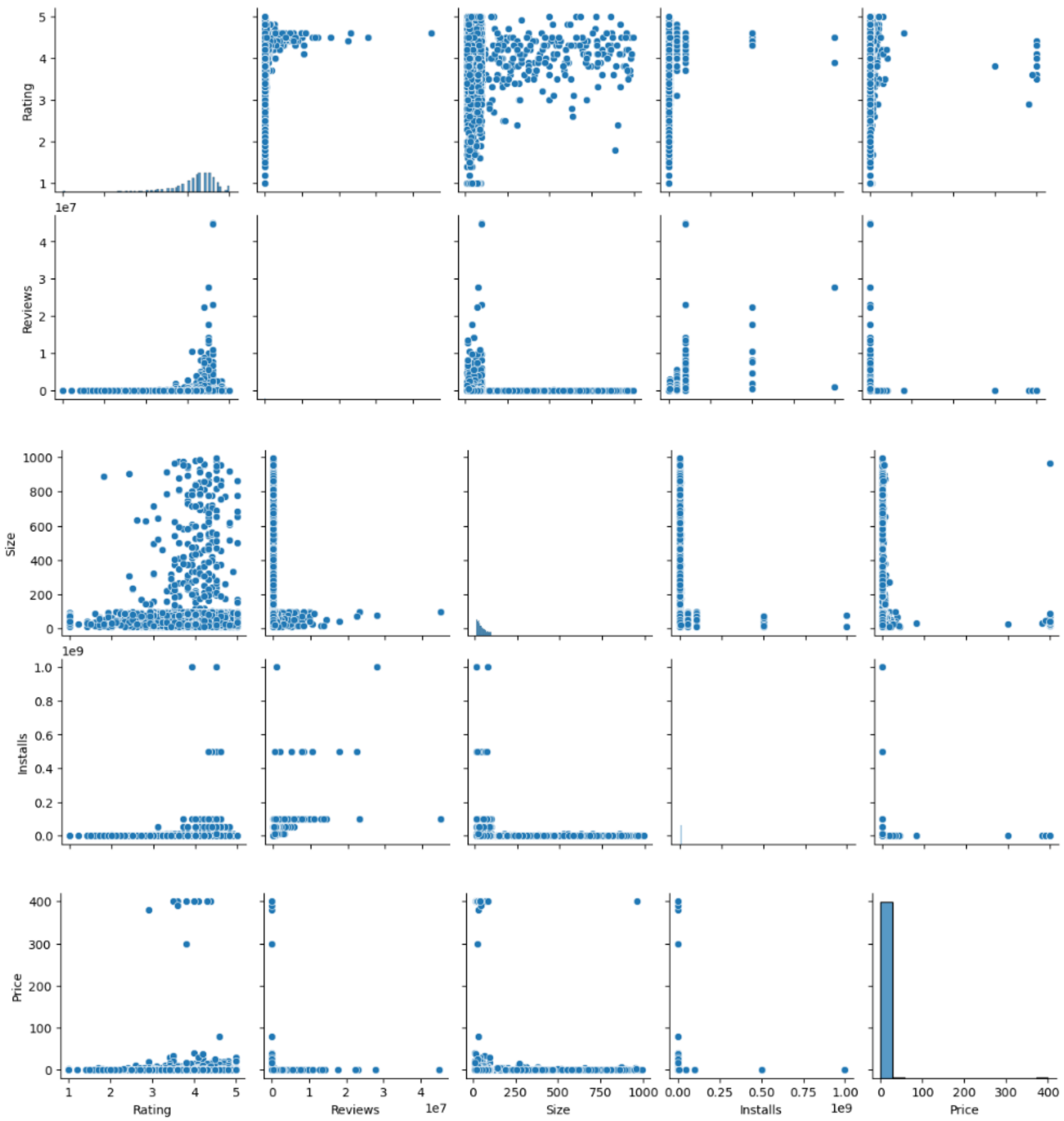
Techniques Analysis

Before applying any clustering technique, we plotted some scatter plots to look at the shape of the data and get an idea of which technique would work better.

1. 3D Scatter plot



2. 2D Pair Scatter Plot

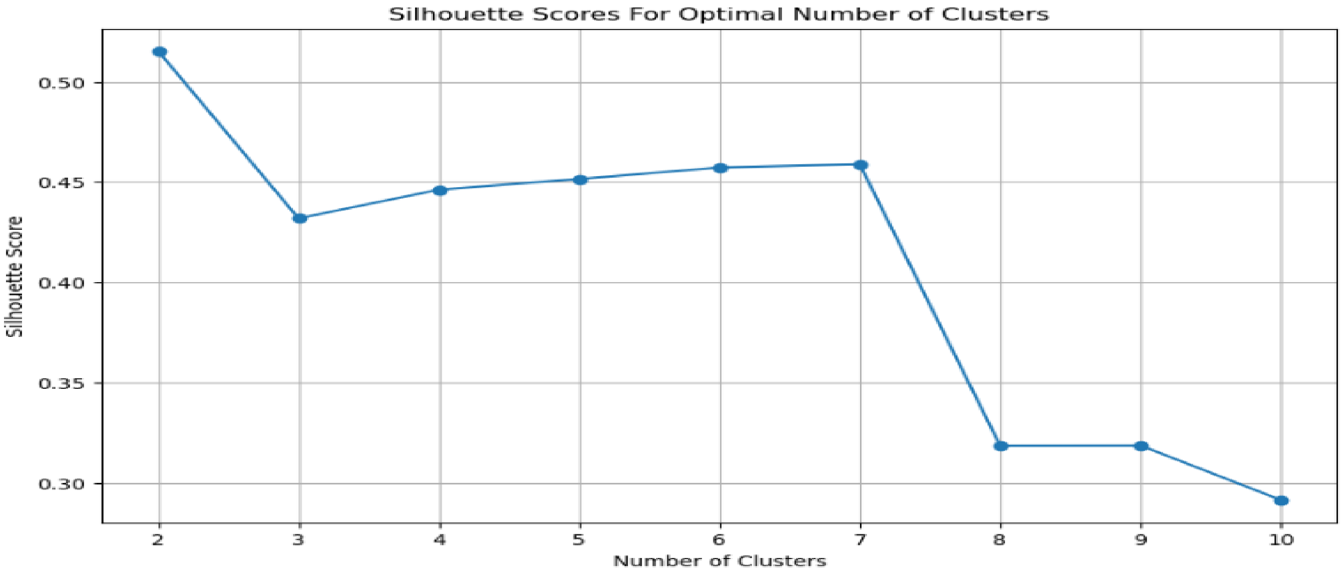


From the 3D scatter plot and 2D scatter pair plot, we can say that, the K-Means Clustering and GMM clustering technique will not give the appropriate result, as K-means assumes clusters in Spherical shape and GMM assumes it in an Elliptical shape, which is not the case here.

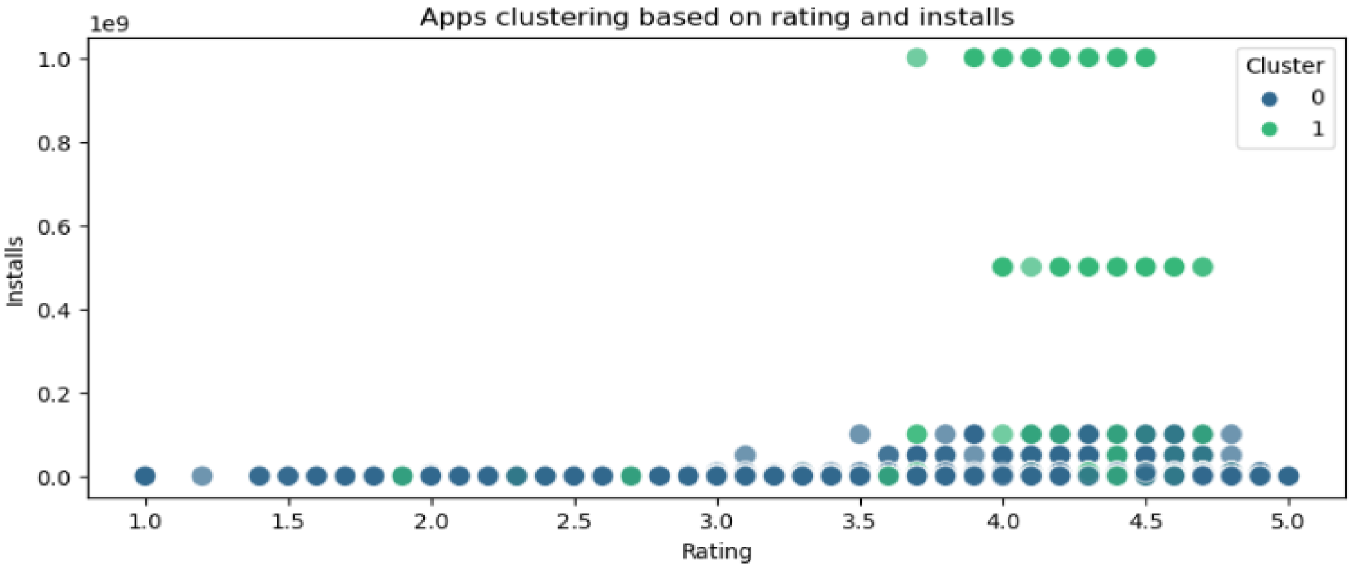
Still, we'll see how these methods perform to confirm our observation.

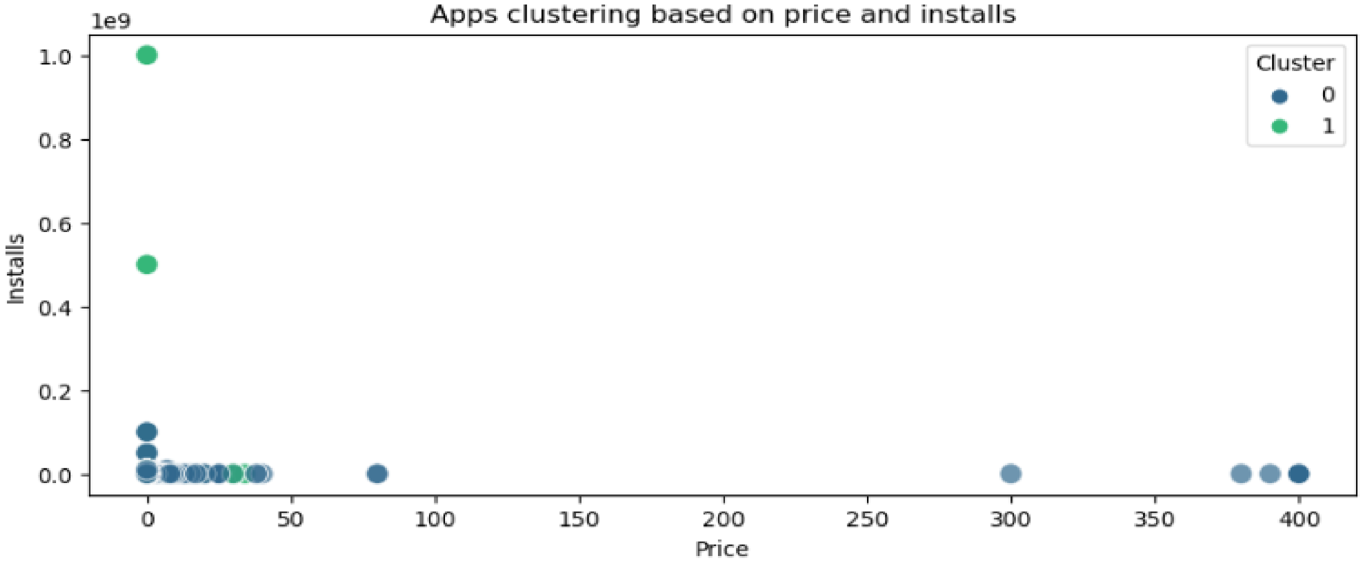
1. K-MEANS CLUSTERING TECHNIQUE:

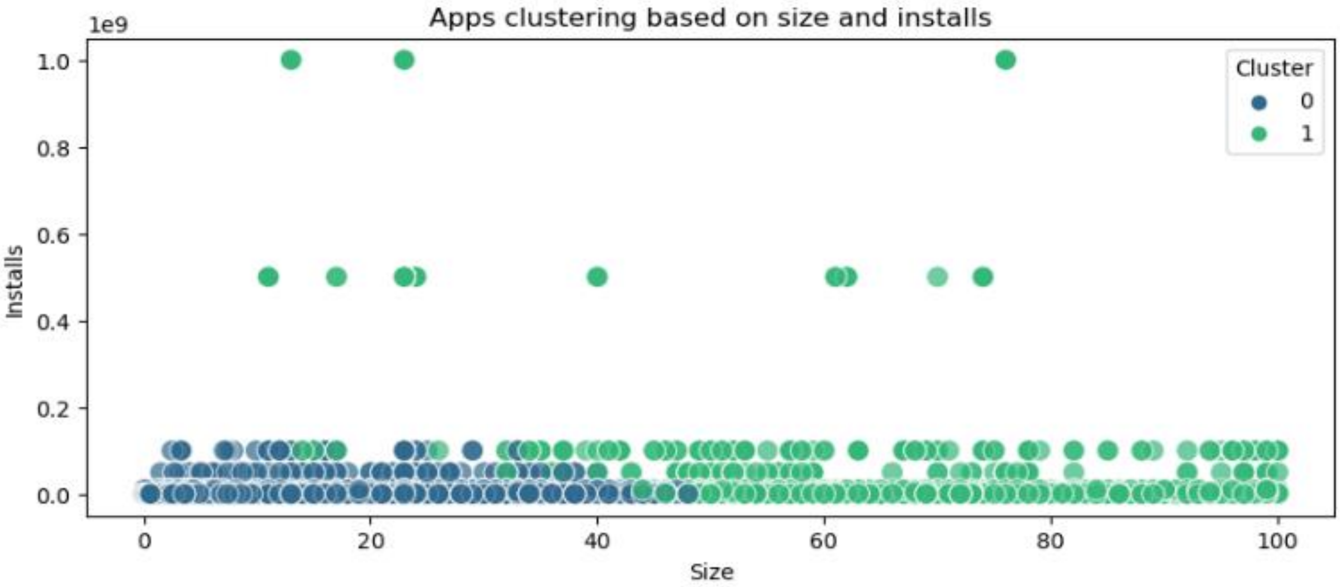
On performing the Silhouette score technique, we got the optimal number of clusters as 2.



which we applied on various features to divide data into 2 clusters.



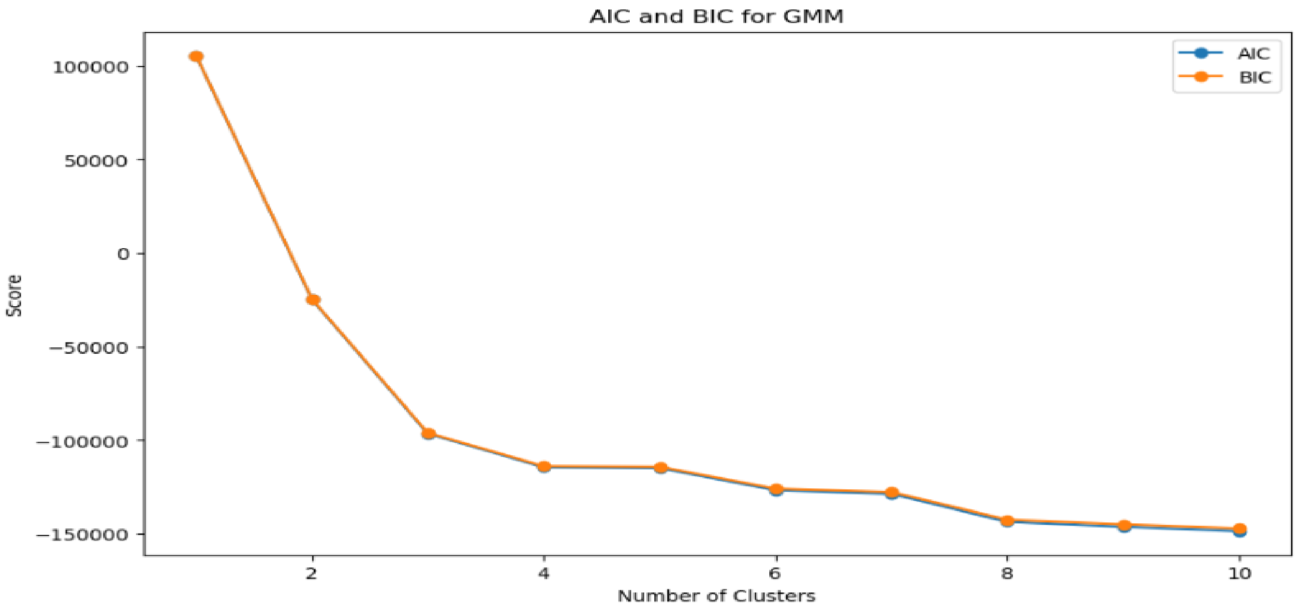




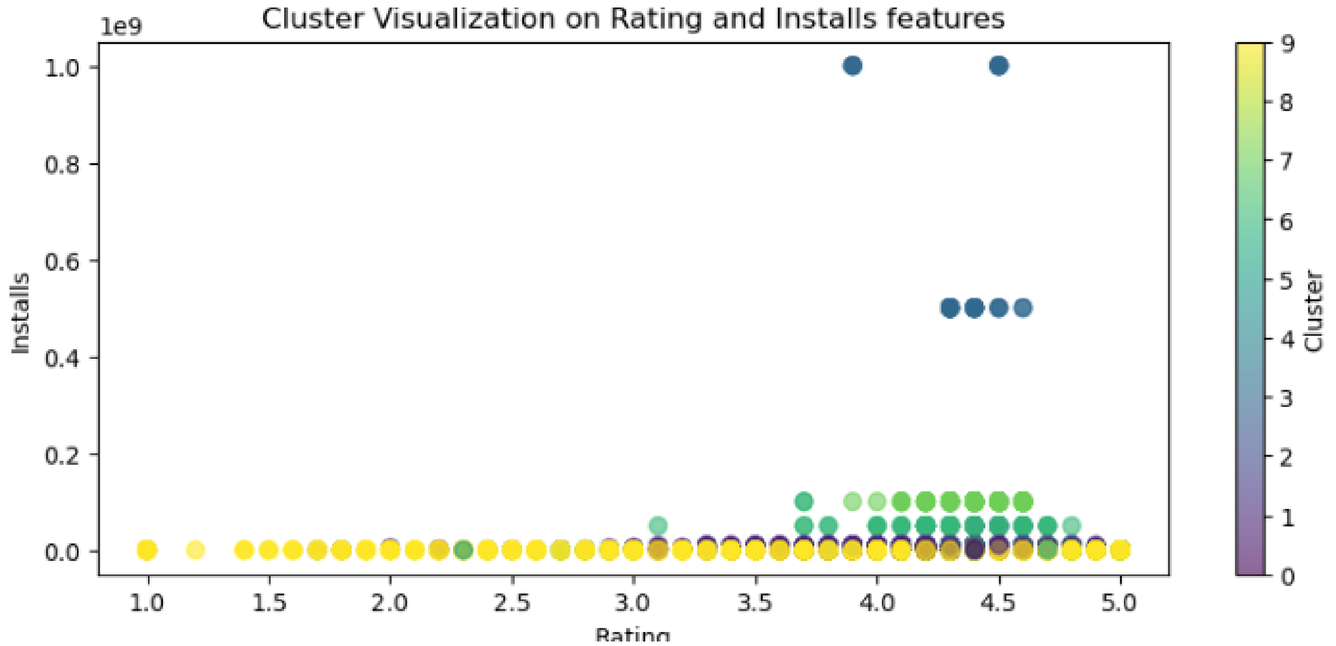
So, when we applied the k-means clustering model on some features, we found that the clusters formed are not very well defined, in all the feature pairs we can see (above) that some data points of one cluster are present in other too.

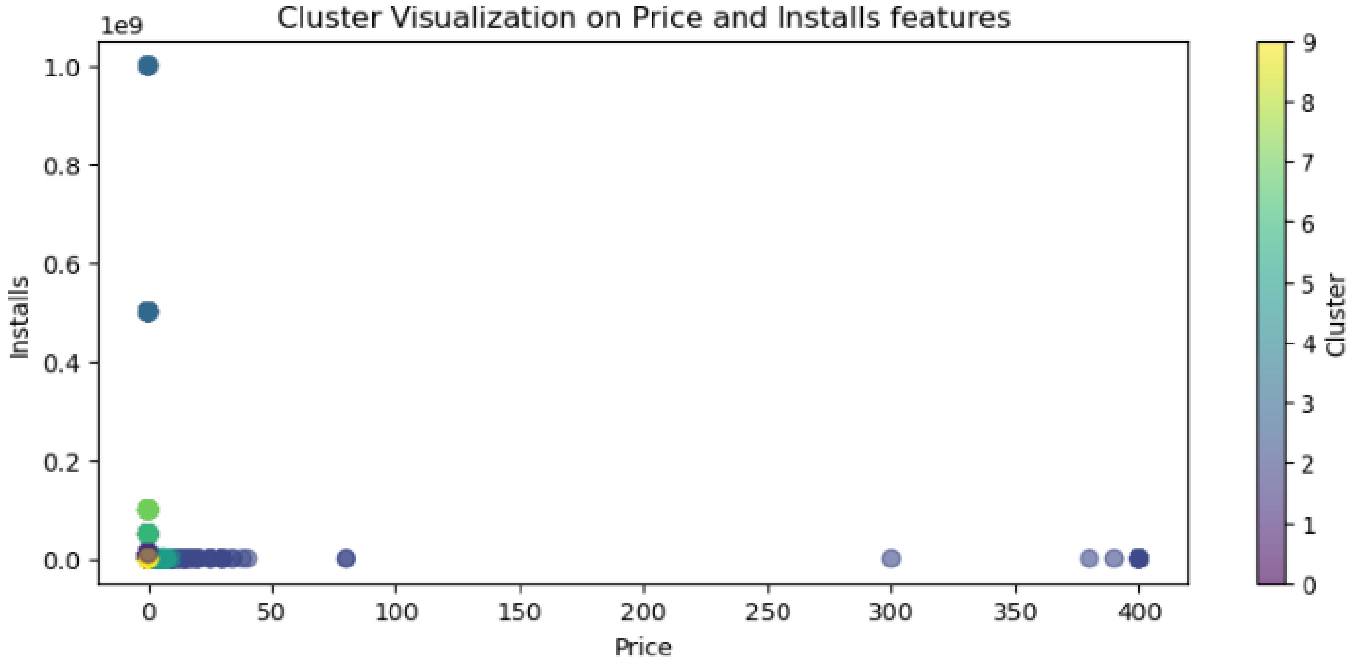
2. GMM (GAUSSIAN MIXTURE MODEL) CLUSTERING TECHNIQUE:

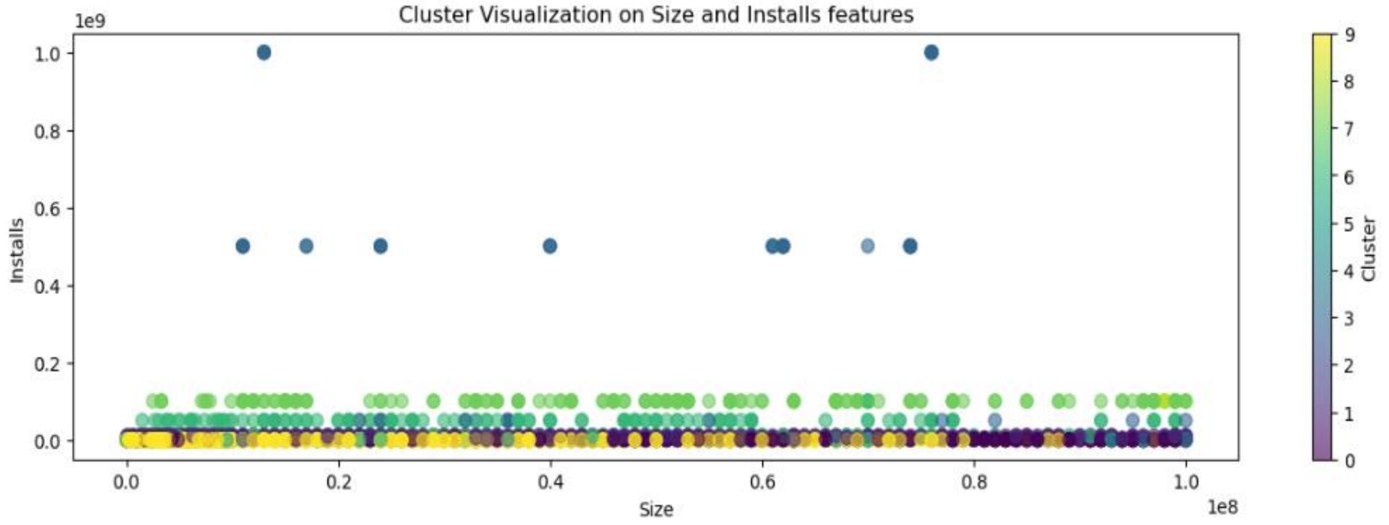
In this technique, we found the optimal number of clusters as 10 by plotting the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) graph, as shown below.



So, when we applied the GMM clustering model on certain features to get 10 clusters, we found that the clusters formed are not well defined and most of them are merging, as shown below.


Cluster Visualization on Rating and Installs features
1e9
1.0
0.8
0.6
0.4
0.2
0.0
Installs
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Rating
Cluster
0
1
2
3
4
5
6
7
8
9

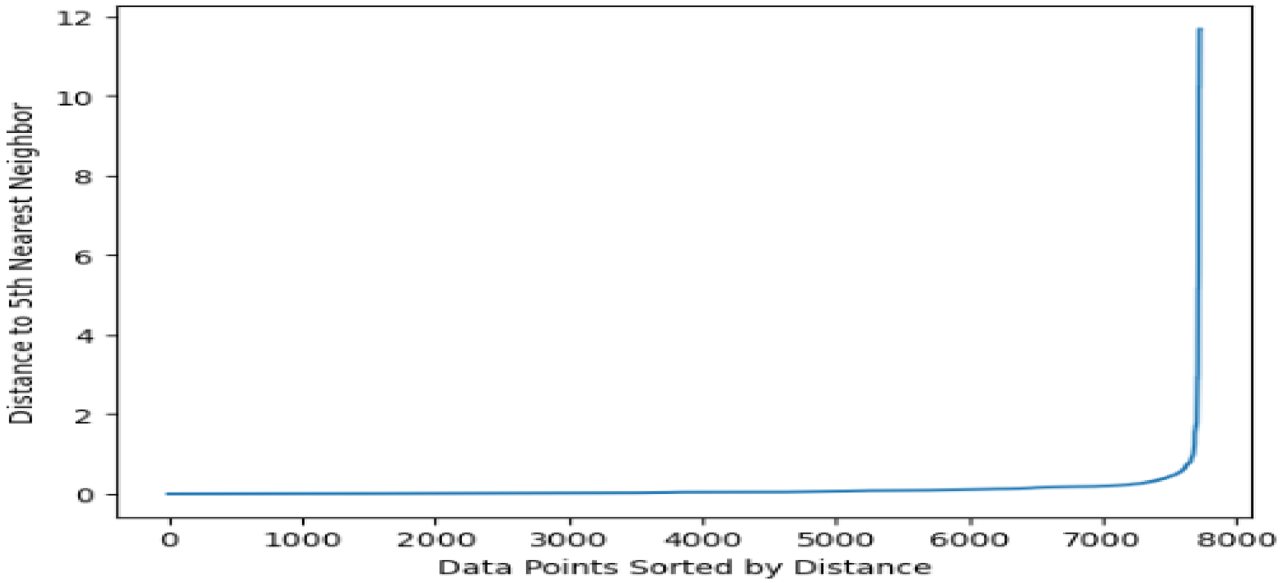

Cluster Visualization on Price and Installs features
1e9
1.0
0.8
0.6
0.4
0.2
0.0
Installs
0
50
100
150
200
250
300
350
400
Price
Cluster
0
1
2
3
4
5
6
7
8
9


Cluster Visualization on Size and Installs features
1e9
1.0
0.8
0.6
0.4
0.2
0.0
Installs
0.0
0.2
0.4
0.6
0.8
1.0
1e8
Size
Cluster
0
1
2
3
4
5
6
7
8
9

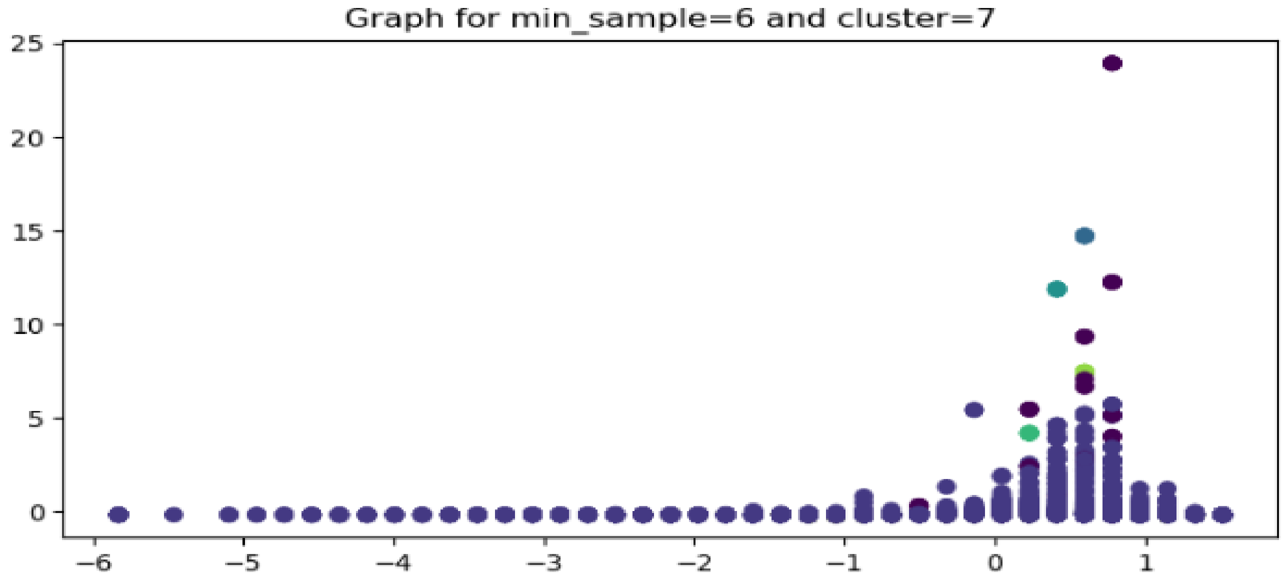
So, we have confirmed that both of these techniques failed to give well-defined clusters for the given data.

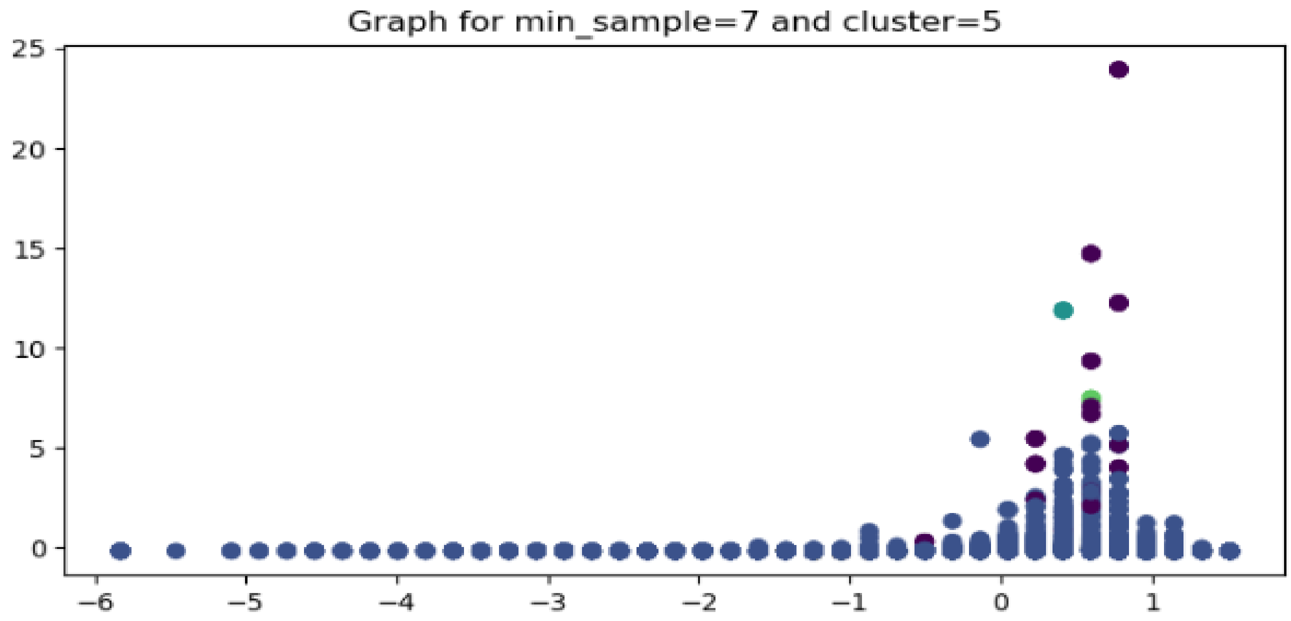
3. DBSCAN (DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE) CLUSTERING TECHNIQUE:

In this technique, we found the optimal value of eps (epsilon) as 1 , by plotting k-nearest neighbour plot, as shown below .

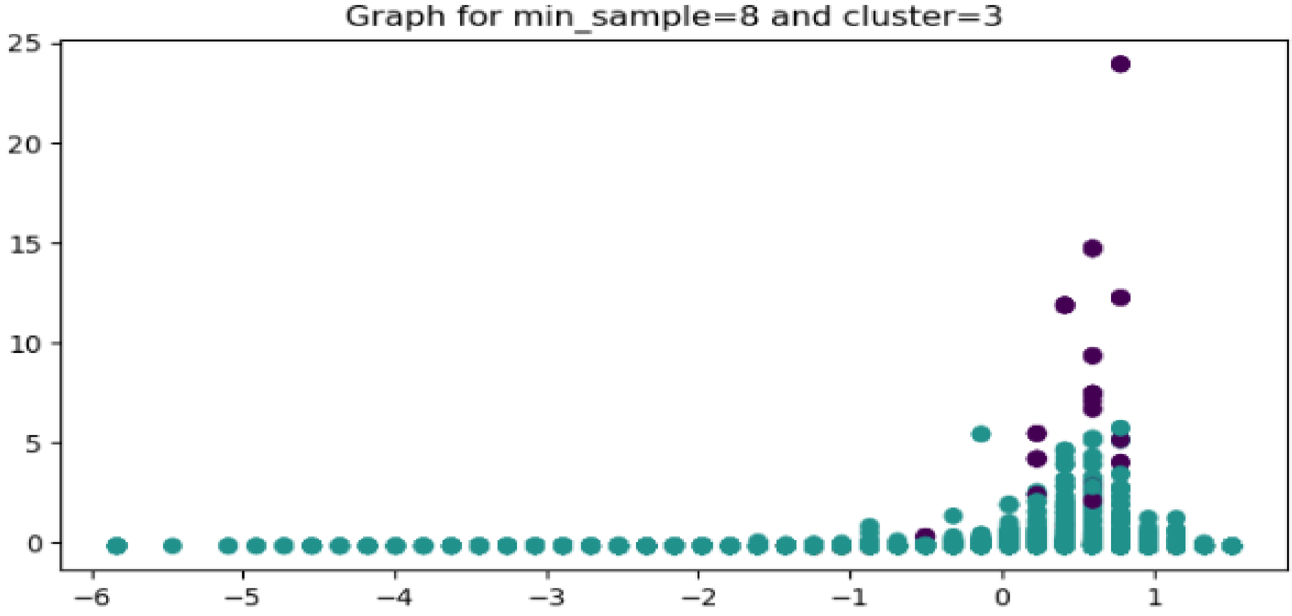


Now, on plotting various graphs with different values of the min_samples parameter we didn't find any number of clusters that grouped the data properly, as shown below.

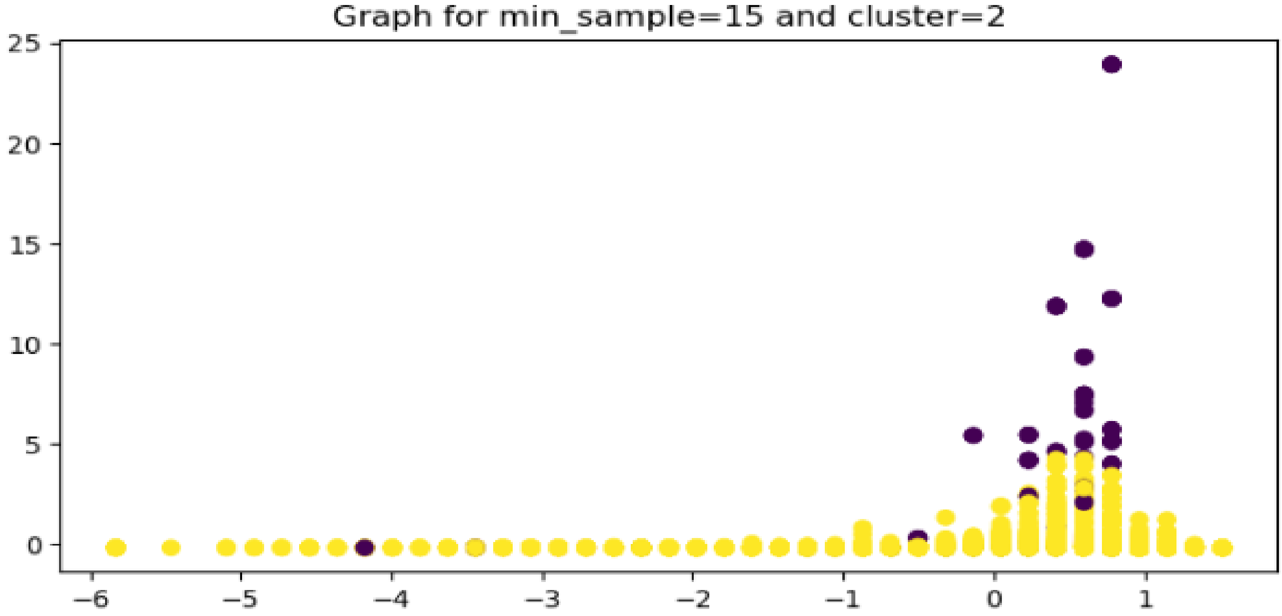




Graph for min_sample=7 and cluster=5



Graph for min_sample=8 and cluster=3

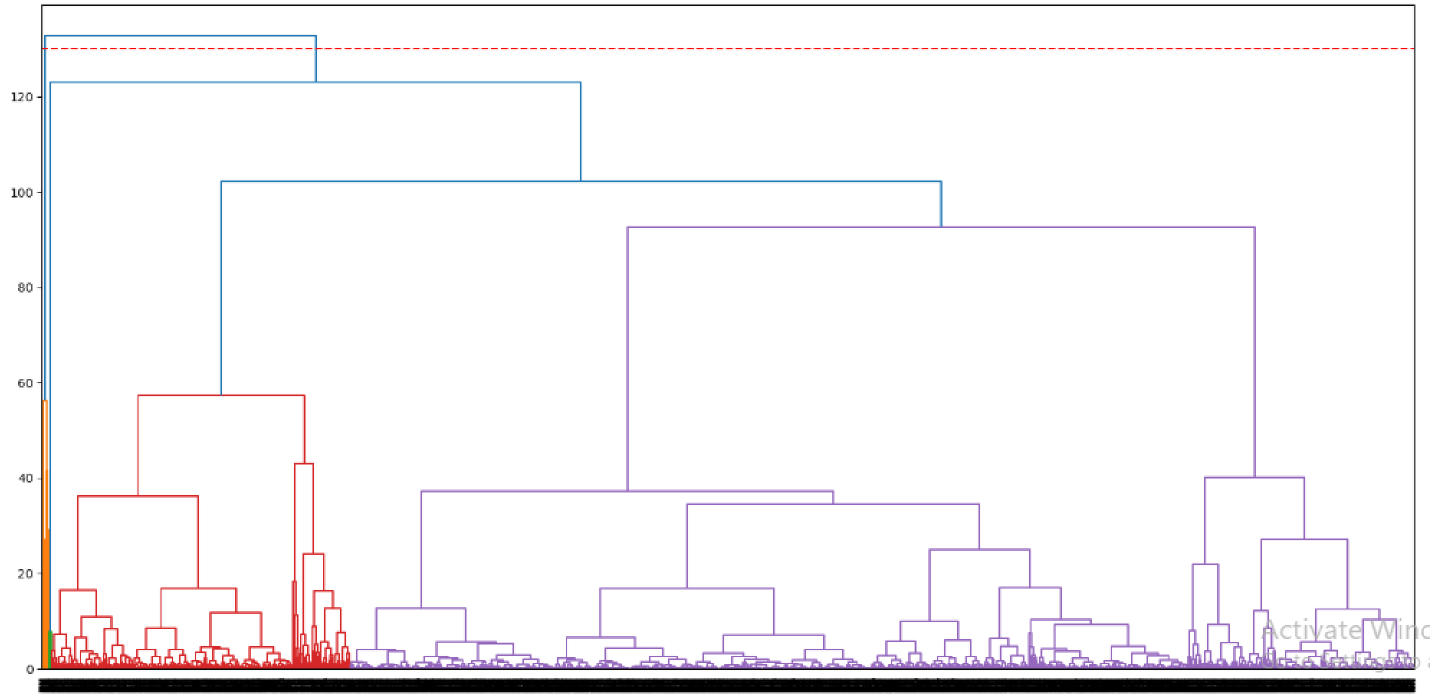


Graph for min_sample=15 and cluster=2

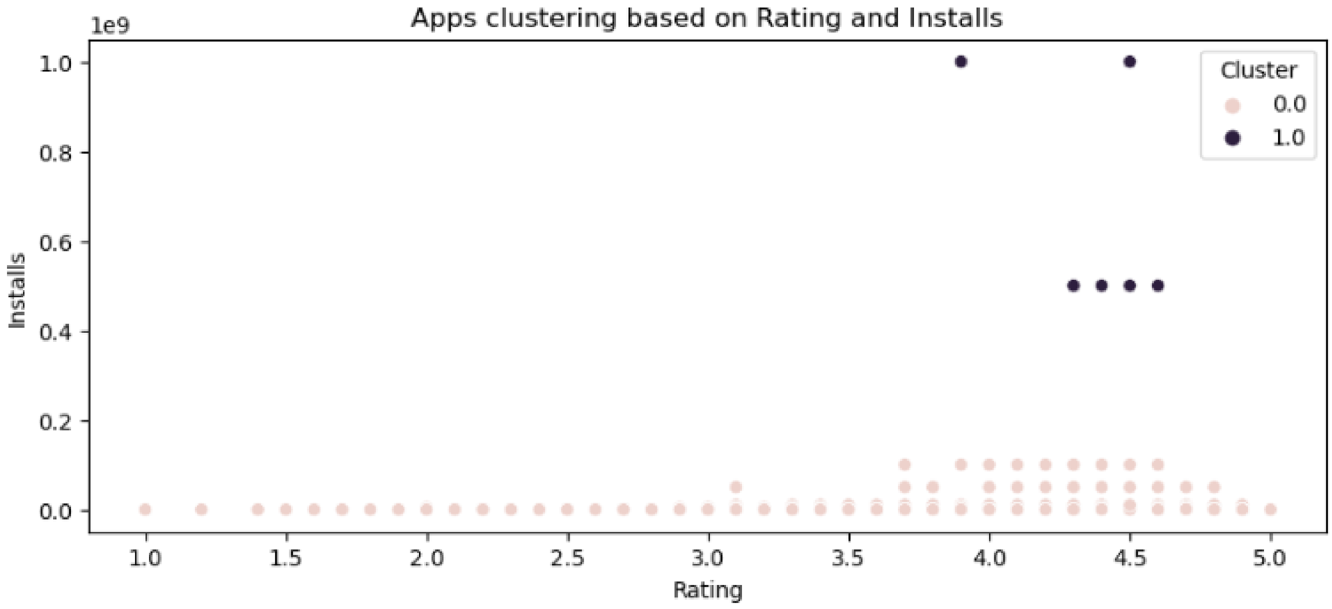
In this clustering method, for any number of clusters above 2, the clusters itself is not visible, and for 2 clusters the data points are not well separated. So even this technique fails to cluster the data effectively.

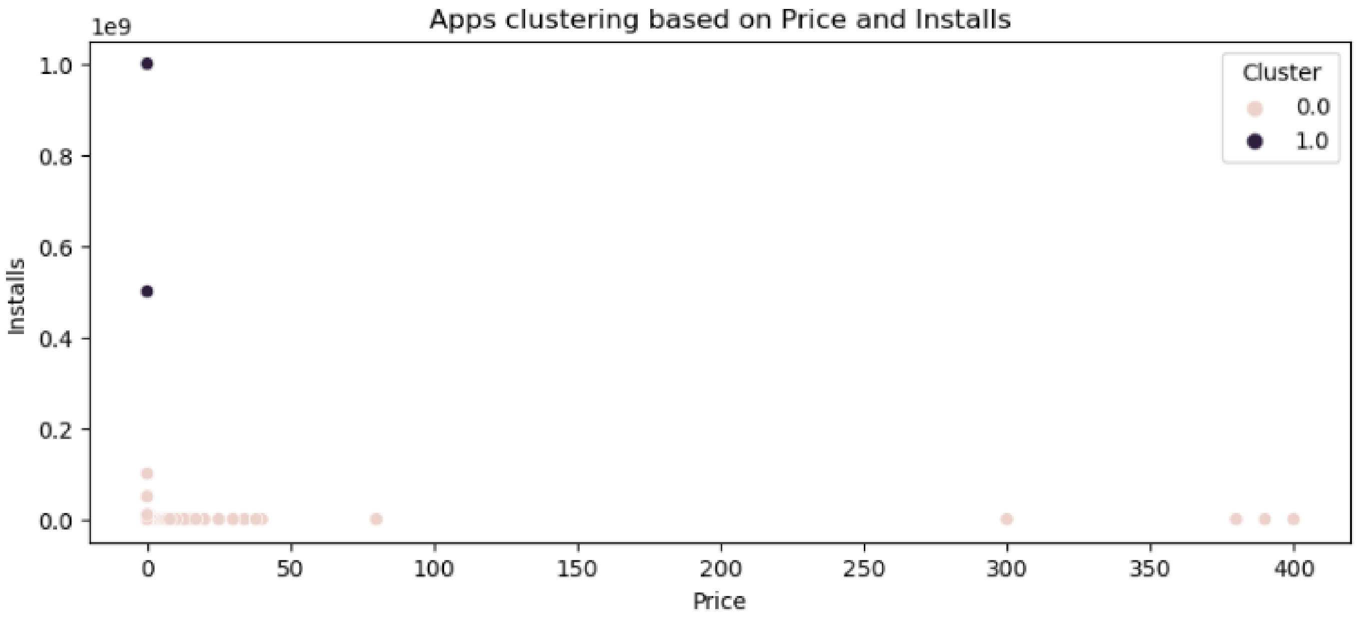
4. HIERARCHICAL CLUSTERING:

In this technique, we have found the optimal number of clusters as 2 by plotting the dendrogram, as shown below.

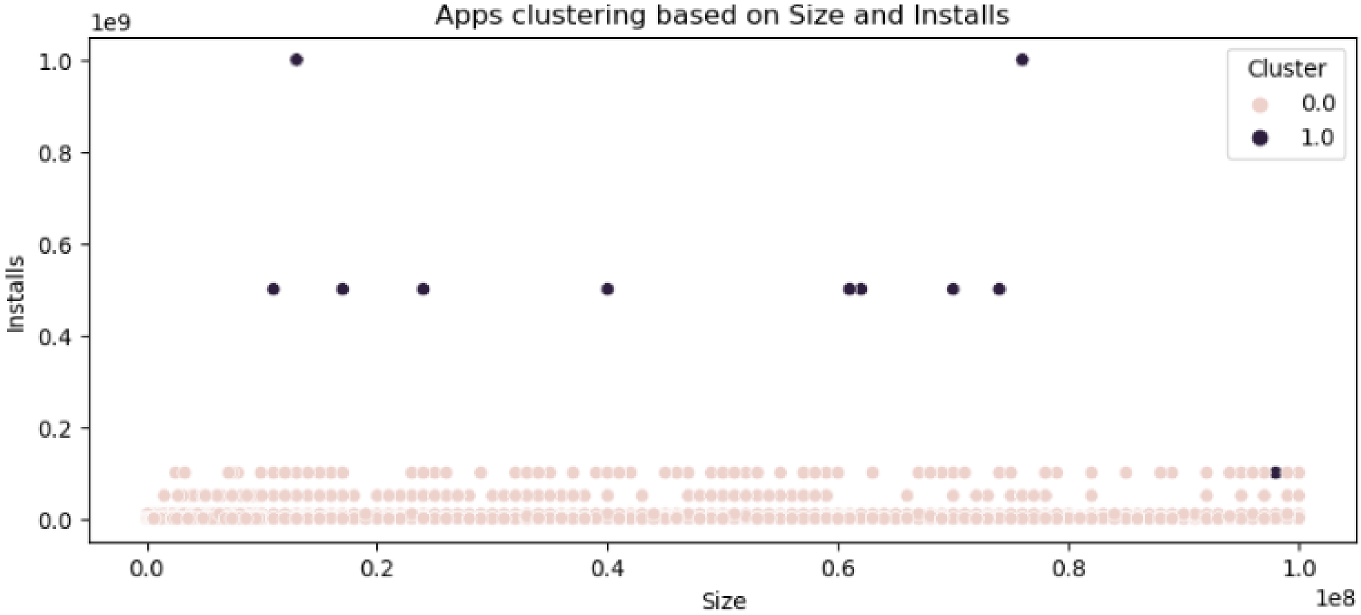


On plotting clusters by applying the Agglomerative Hierarchical Clustering model with various features, we found the clusters formed are well defined and all the data points are also well allocated into different clusters, as shown below.

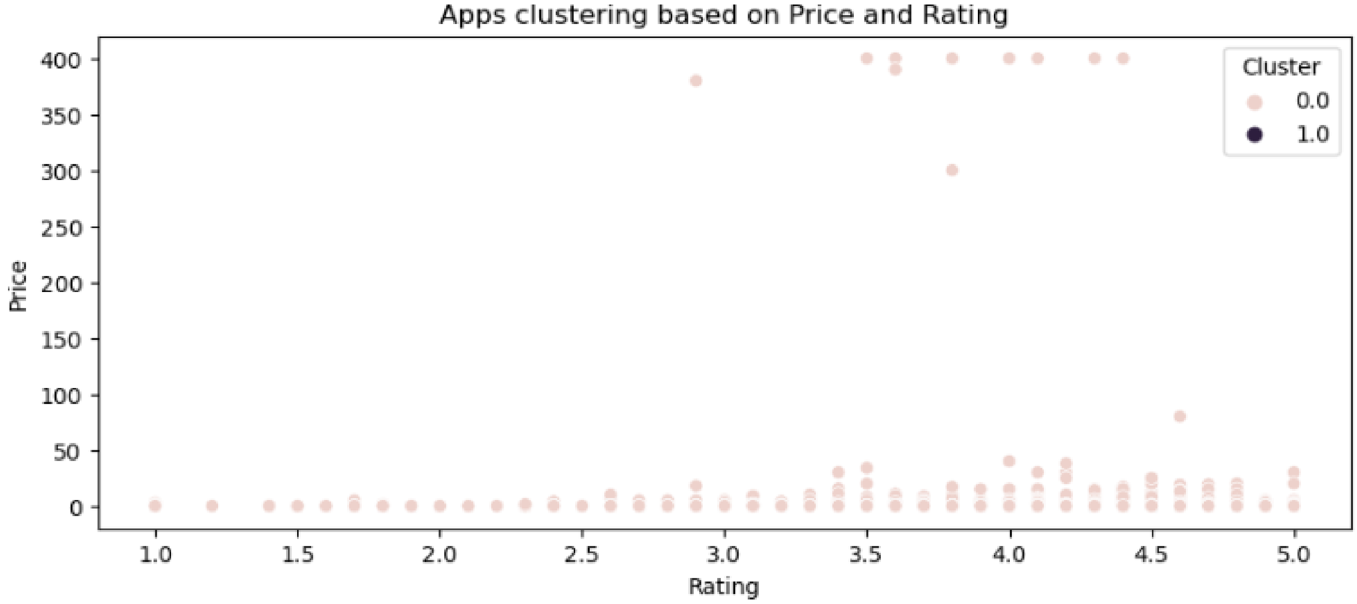




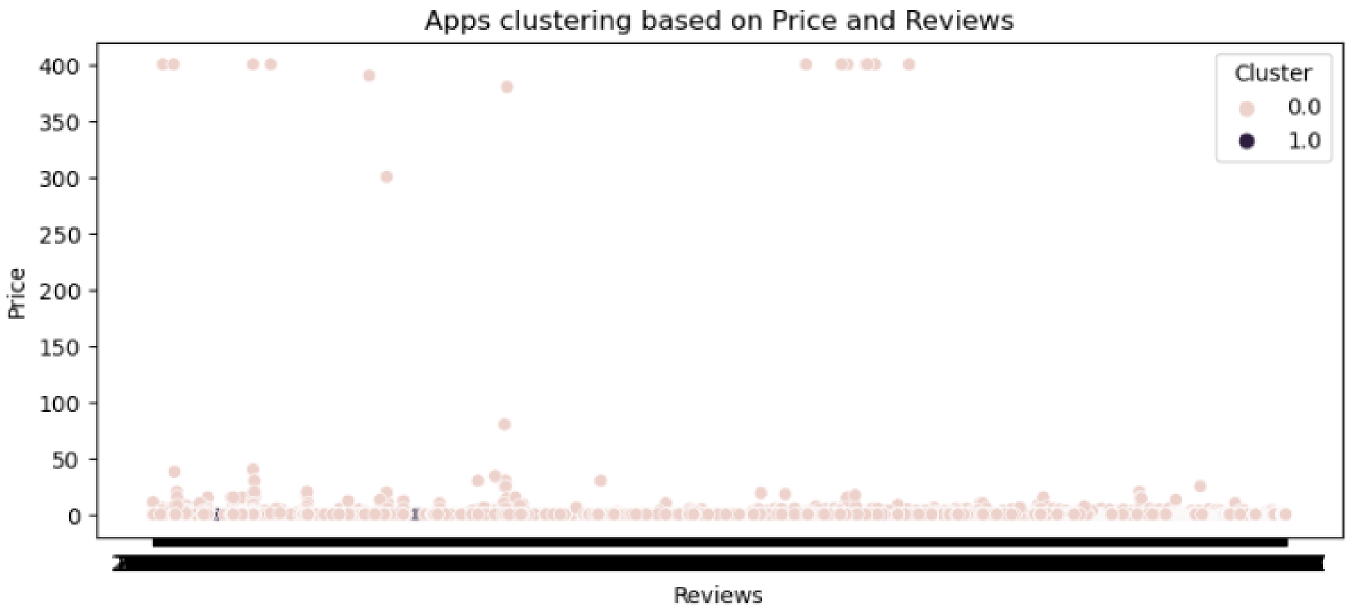
Apps clustering based on Price and Installs

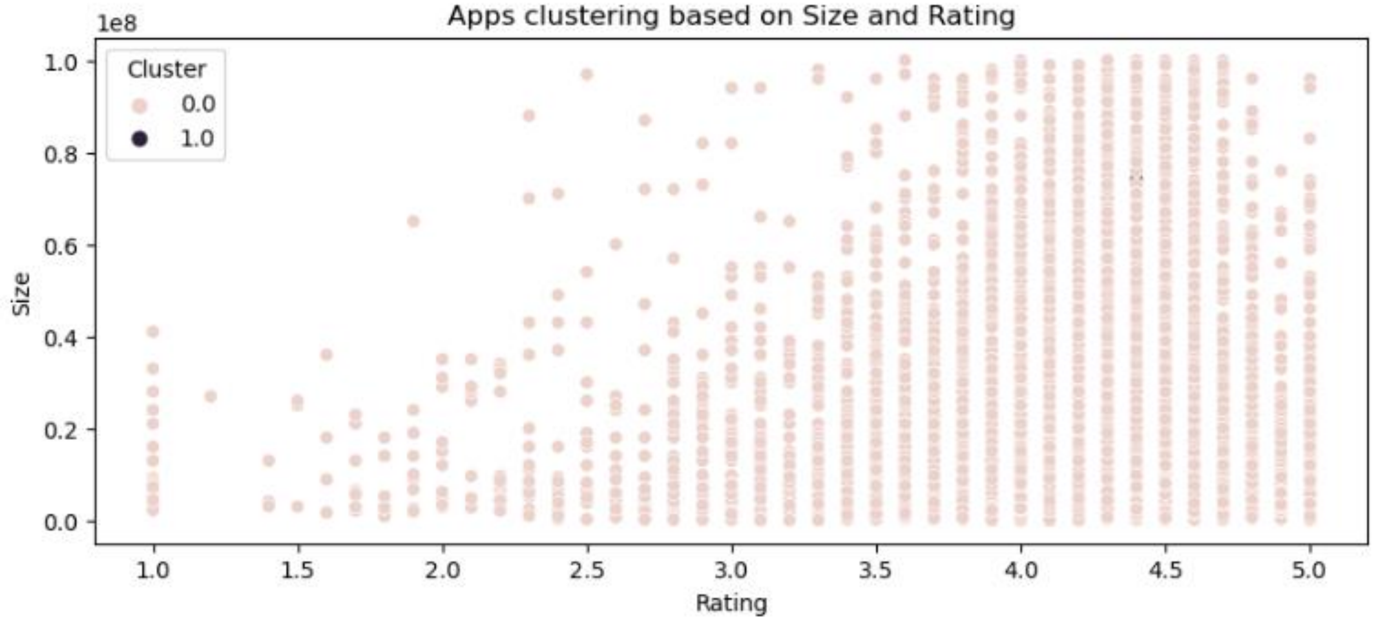


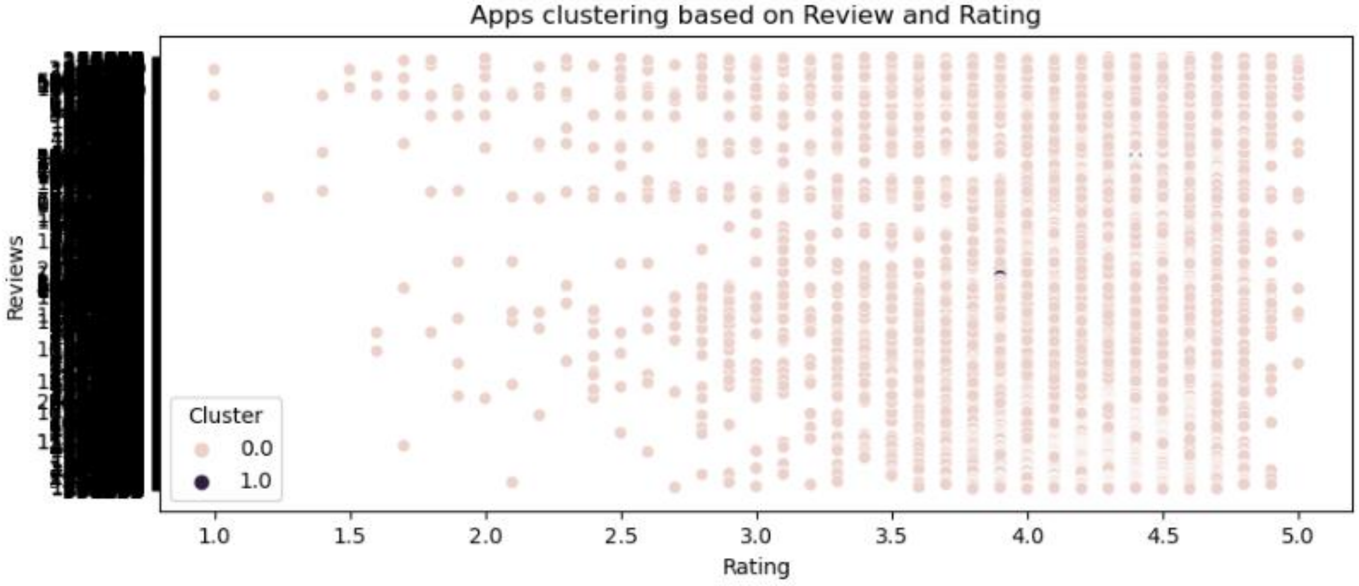
Apps clustering based on Size and Installs



Apps clustering based on Price and Rating


Apps clustering based on Price and Reviews
Cluster
0.0
1.0
Price
400
350
300
250
200
150
100
50
0
Reviews


Apps clustering based on Size and Rating
1e8
Cluster
0.0
1.0
Size
1.0
0.8
0.6
0.4
0.2
0.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Rating


Apps clustering based on Review and Rating
Cluster
0.0
1.0
Reviews
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Rating

Conclusion

Hence, we can conclude that the Agglomerative Hierarchical Clustering technique is the most suitable to cluster the Google Play Store data for analysis of applications, as in the other techniques like K-Means, GMM, and DBSCAN clusters are not formed in a well-defined manner due to which applications can't be grouped properly for effective analysis.

Analysis of Google Play store Dataset based on Agglomerative Hierarchical Clustering

The following insights are based on the clustering analysis of the Google Play Store dataset using Hierarchical Clustering. The analysis was performed to understand relationships between key variables such as installs, ratings, prices, app sizes, and reviews which in turn will help app developers to further improve.

Key Insights:

1. Installs vs. Rating -

- **Insight:** Apps with **lower ratings** tend to have **lower installs**. However, apps with **higher ratings** can have both **low** and **high installs**.

- **Interpretation:** A high rating does not always guarantee more installs. Other factors like app visibility, marketing, and user preferences might influence installs even if the app has a high rating.

- **Recommendation:** App developers should focus on **user engagement and marketing efforts** to increase installs, especially for highly-rated apps.

2. Installs vs. Pricing -

- **Insight:** Apps with a **lower price** or are **free** tend to have **higher installs**, while apps with **higher prices** generally have **lower installs**. There is no instance of an app with both **high prices** and **installs.**

- **Interpretation:** Price sensitivity is evident in the app market, with users favoring free or low-cost apps. High-priced apps may struggle to gain large user bases.

- **Recommendation:** Developers should consider **freemium models** or **discount strategies** to encourage more downloads, especially for premium apps.

3. App Size vs. Installs -

- **Insight:** App size does not appear to significantly affect the number of installs. Both large and small apps exhibit a similar pattern of installs.

- **Interpretation:** Users may prioritize app functionality and performance over size, suggesting that size alone is not a determining factor for installs.

- **Recommendation:** While optimizing app size for smoother downloads is important, **focus on app features and performance** is likely to have a greater impact on installs.

4. Rating vs. Pricing -

- **Insight: Paid apps** generally have **better ratings**, ranging between 2.9 and 4.5. In contrast, **free or cheaper apps** (priced up to Rs. 100) have ratings spread across the entire scale.

- **Interpretation:** Users may have **higher expectations** for paid apps, leading to more positive reviews, while free apps may attract a wider variety of user experiences and ratings.

- **Recommendation:** To maintain high ratings for paid apps, developers should focus on **premium quality**, **consistent updates**, and **customer support** to meet users' higher expectations.

5. Rating vs. Reviews -

- **Insight:** Apps with **higher ratings** tend to receive **more reviews**.

- **Interpretation:** Users are generally more inclined to leave reviews when they have a **positive experience** with the app, leading to a correlation between high ratings and review counts.

- **Recommendation:** Encourage user reviews by offering **in-app incentives** for feedback both positive and negative, and work to rectify the issue by seeing the negative feedback and turning it into a positive one. This can enhance both app visibility and credibility.

6. Price vs. Reviews -

- **Insight: High-priced apps** have **fewer reviews**, while **free or cheap apps** have **more reviews**, likely due to the higher number of installs.

- **Interpretation:** The **barrier to entry** for free or cheap apps is lower, leading to a broader user base and more reviews. In contrast, higher-priced apps might have fewer users, resulting in fewer reviews.

- **Recommendation:** Developers of high-priced apps should encourage reviews from their niche user base by providing **exceptional value and service** and promoting **review requests** after positive user experiences.

7. App Size vs. Rating -

- **Insight:** There is no clear relationship or pattern between **app size** and **rating**.

- **Interpretation:** App size does not seem to influence user ratings, indicating that users rate apps based on factors such as usability, performance, and overall experience rather than size.

- Recommendation: Developers should prioritize improving **app quality and user experience**, rather than focusing solely on size optimization. However, **performance** should be regularly monitored to avoid issues related to larger apps, such as sluggish performance on lower-end devices.

General Recommendations:

Based on these insights, several strategic recommendations emerge for app developers and businesses looking to optimize their app's success:

1. Focus on User Experience: High ratings are associated with better installs and reviews, making ****user satisfaction**** critical for success. Regular updates, bug fixes, and new features should be prioritized.

2. Adopt a Freemium Pricing Model: To appeal to a wider user base and increase installs, consider adopting a **freemium** model, where basic features are free, and premium features require payment. This can help bridge the gap between free app popularity and revenue generation.

3. Encourage User Reviews: Actively prompting users to leave reviews, especially after positive interactions, can increase the number of reviews and boost visibility on the app store.

4. Optimize Marketing for High-Rated Apps: While high ratings are important, they must be accompanied by **effective marketing strategies** to convert them into higher install numbers.

5. Price Sensitivity: For high-priced apps, offer **trial versions** or **promotional discounts** to attract a broader audience and increase the number of reviews, which will build trust with potential buyers.

By following these insights and recommendations, developers can maximize their app's potential, improve user satisfaction, and ultimately drive greater success in the competitive app market.