Google Play

Unsupervised Machine Learning Project

Dataset

~10 800 instances after cleaning

And there was a lot of cleaning. Started out with 1 numerical col

Columns: App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating

5 clusters using Kmeans & their characteristics

0: 1770 apps, all free, cluster has 2nd most downloads and reviews, largest mean size of its apps and all categories are represented

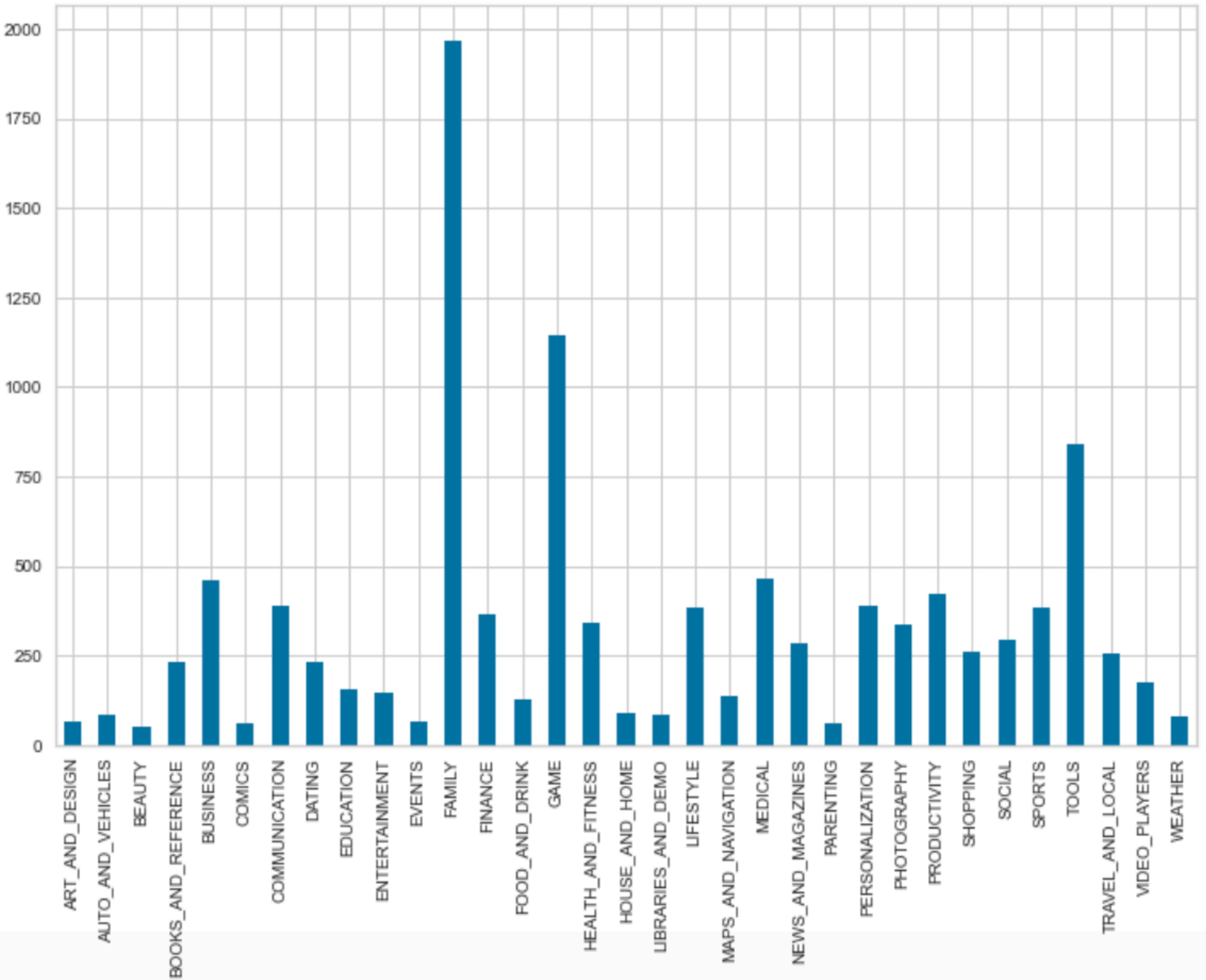
1: 115 apps, all free, cluster has by far the largest avg installs and avg reviews but also total installs, 2nd largest mean app size and only 12 categories represented

2: 18 apps, all paid for (between 200-400$), tad bit lower average rating and only 4 categories represented

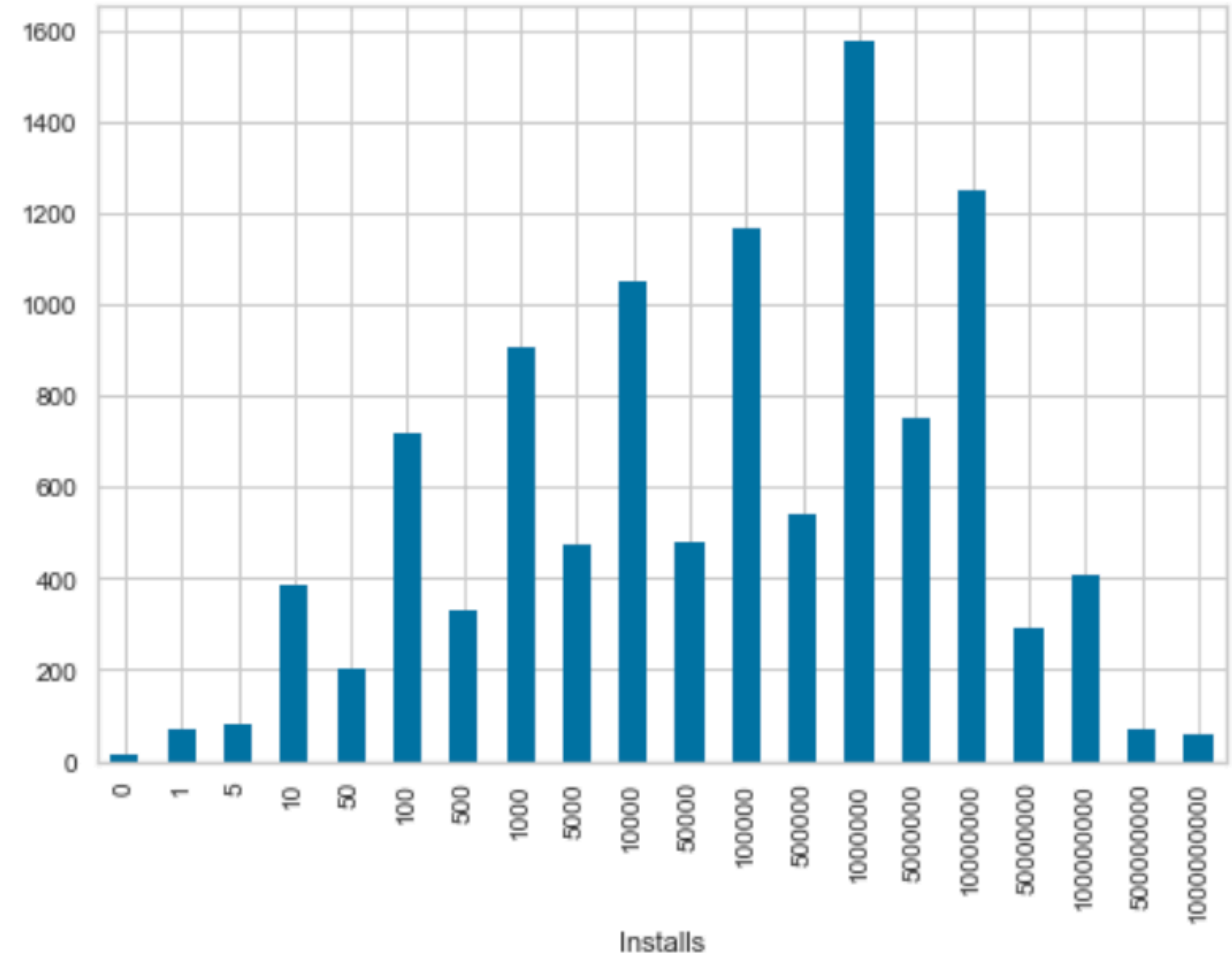
3: 770 apps, all paid for (under 200$), very few total installs, almost all categories represented (30/33)

4: 8147 apps, all free, all cats represented

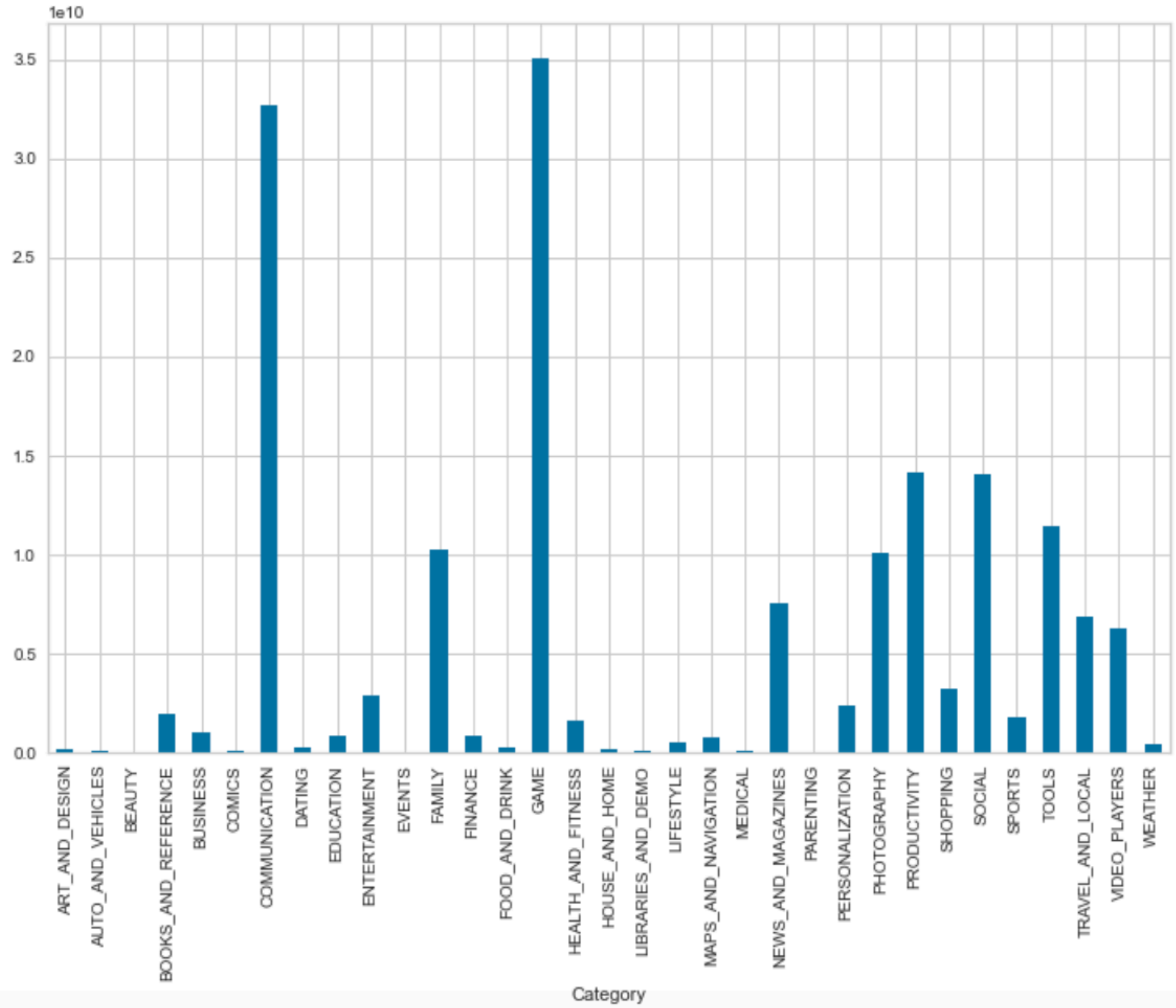
How the apps are distributed over the Categories



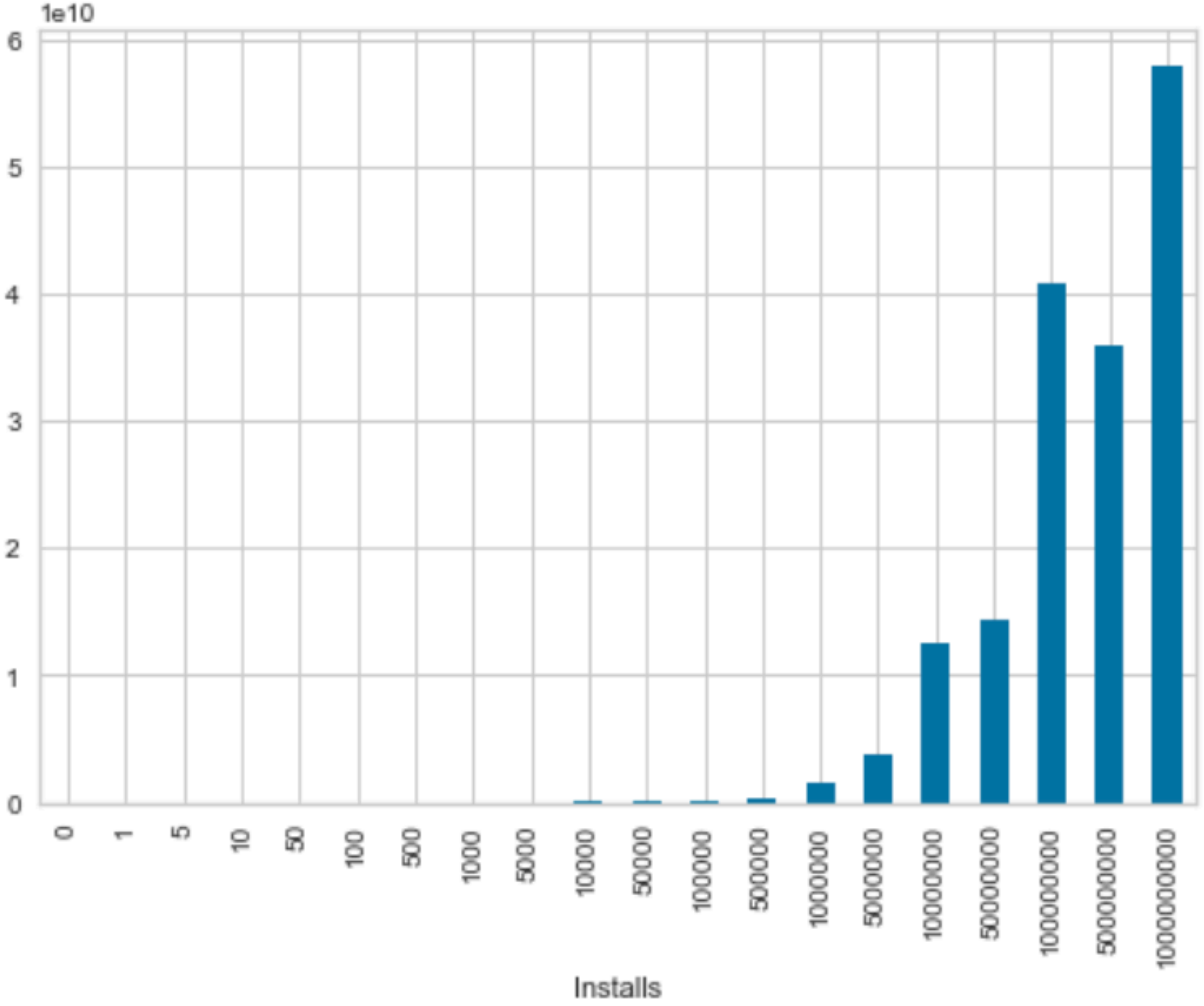
Distribution of downloads



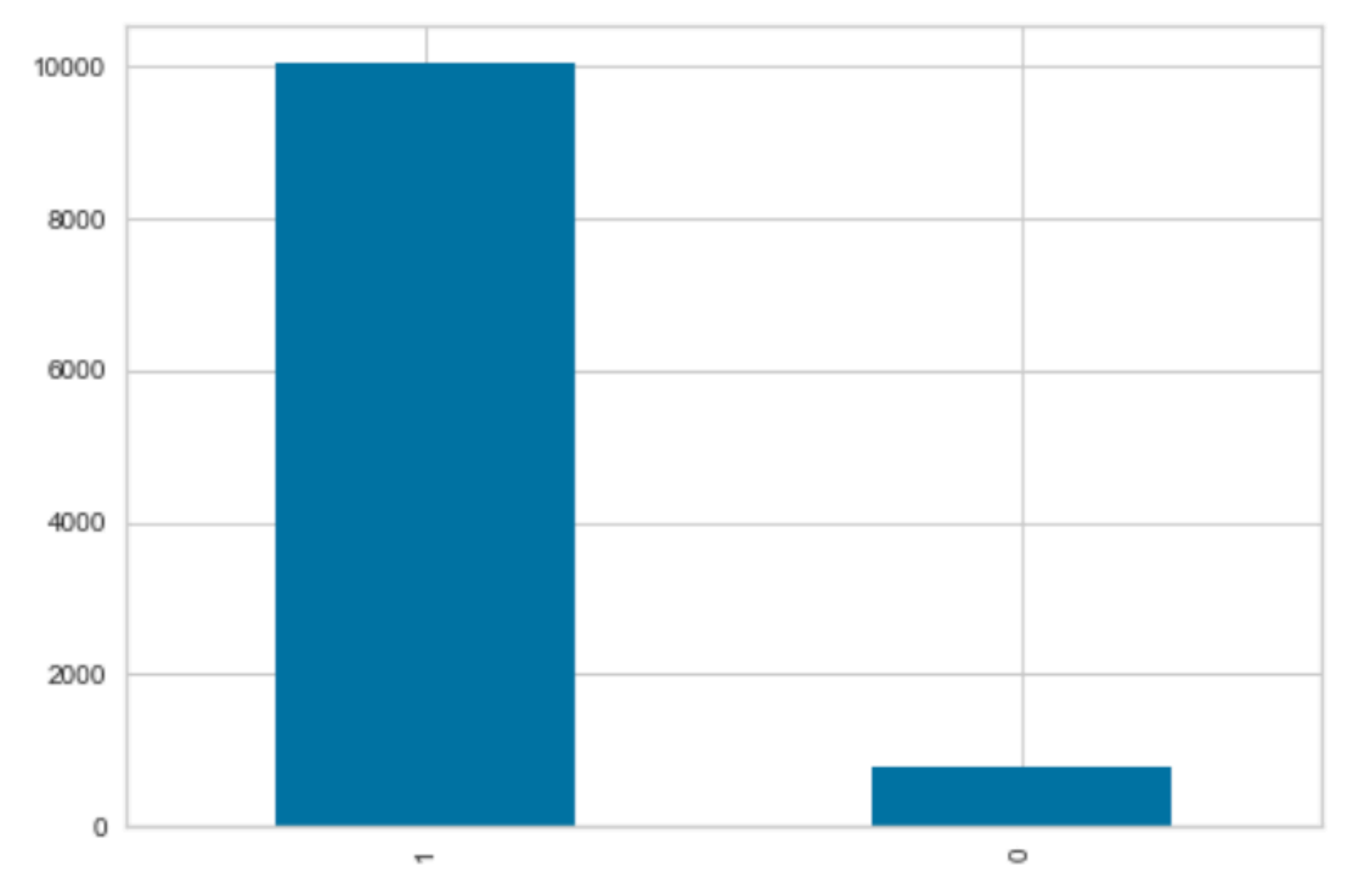
Number of downloads of apps in each category



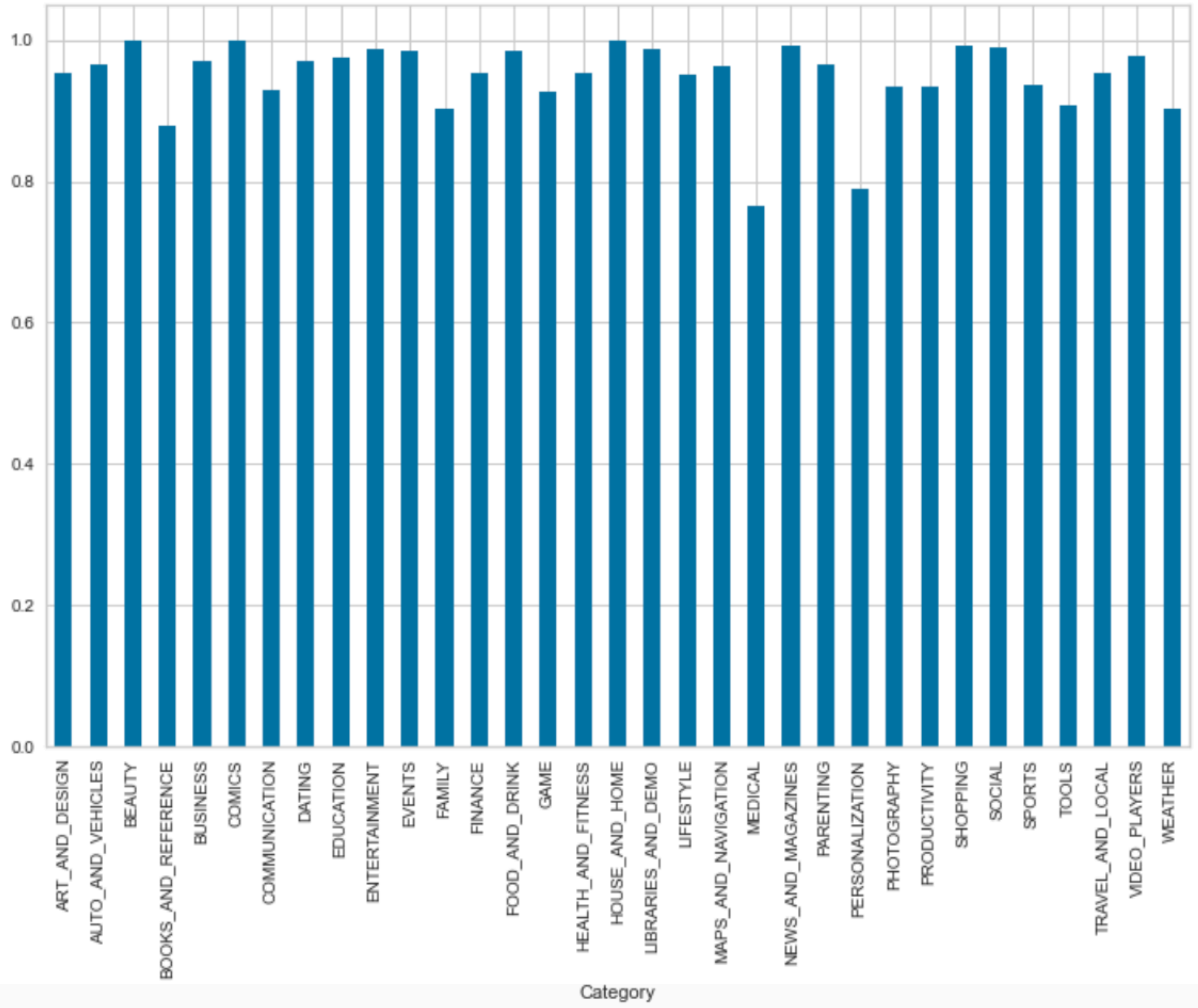
Distribution of total downloads (167 631 856 377)



- 1 = Free app
- 0 = Paid app

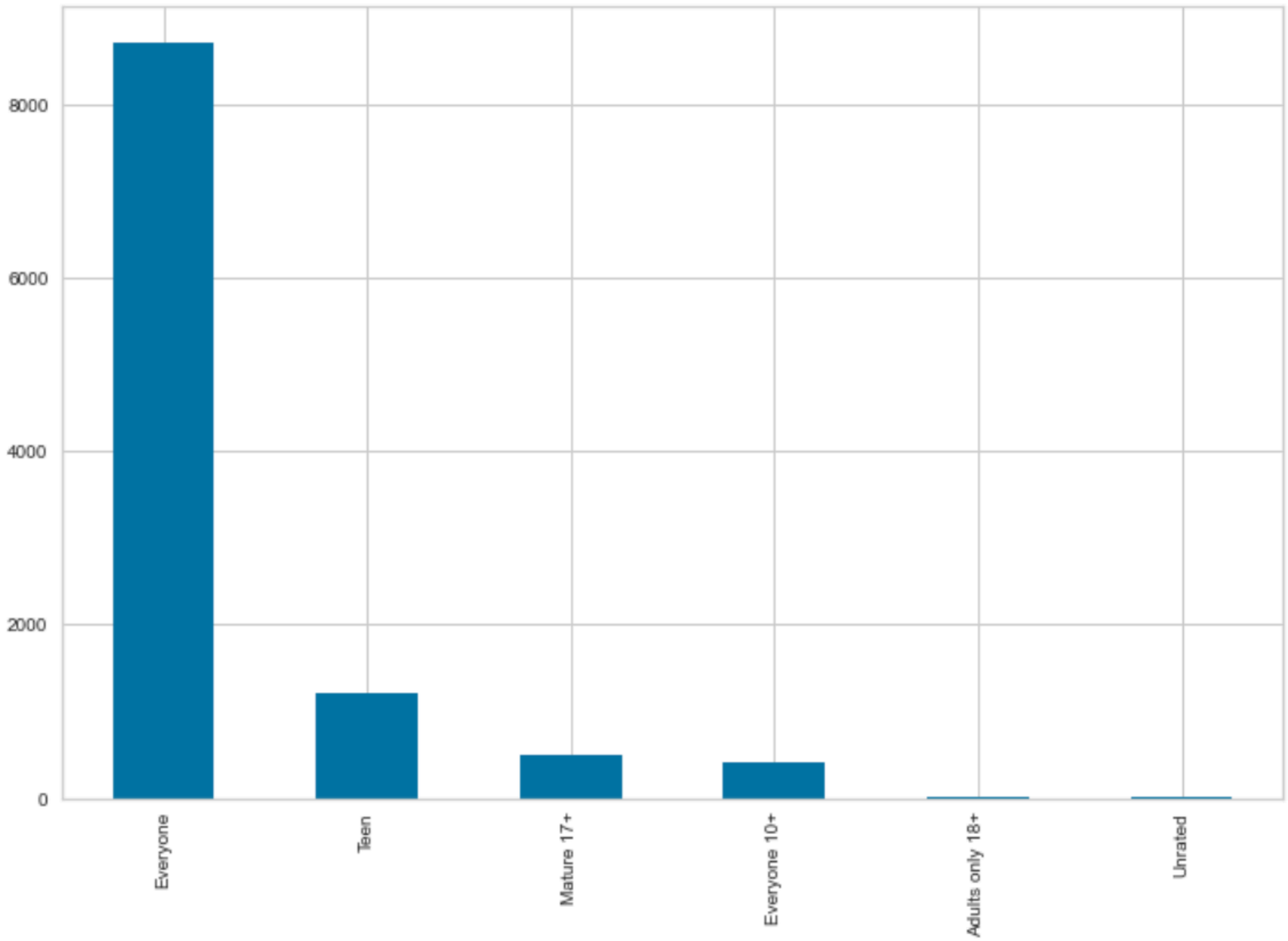


How big a percentage of apps in each Category are Free



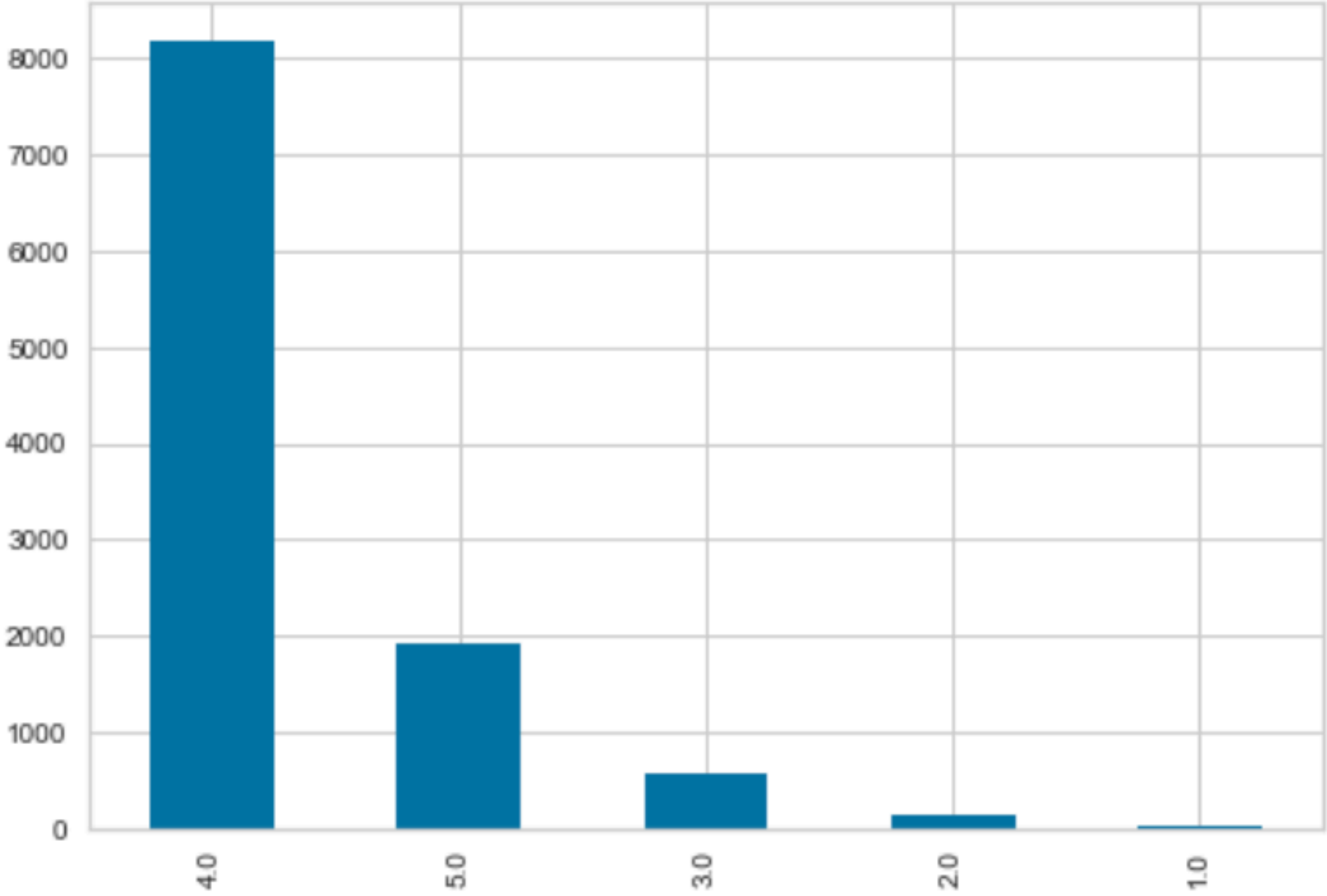
Content Rating distribution

Very few are rated 18+ and a vast majority is for Everyone

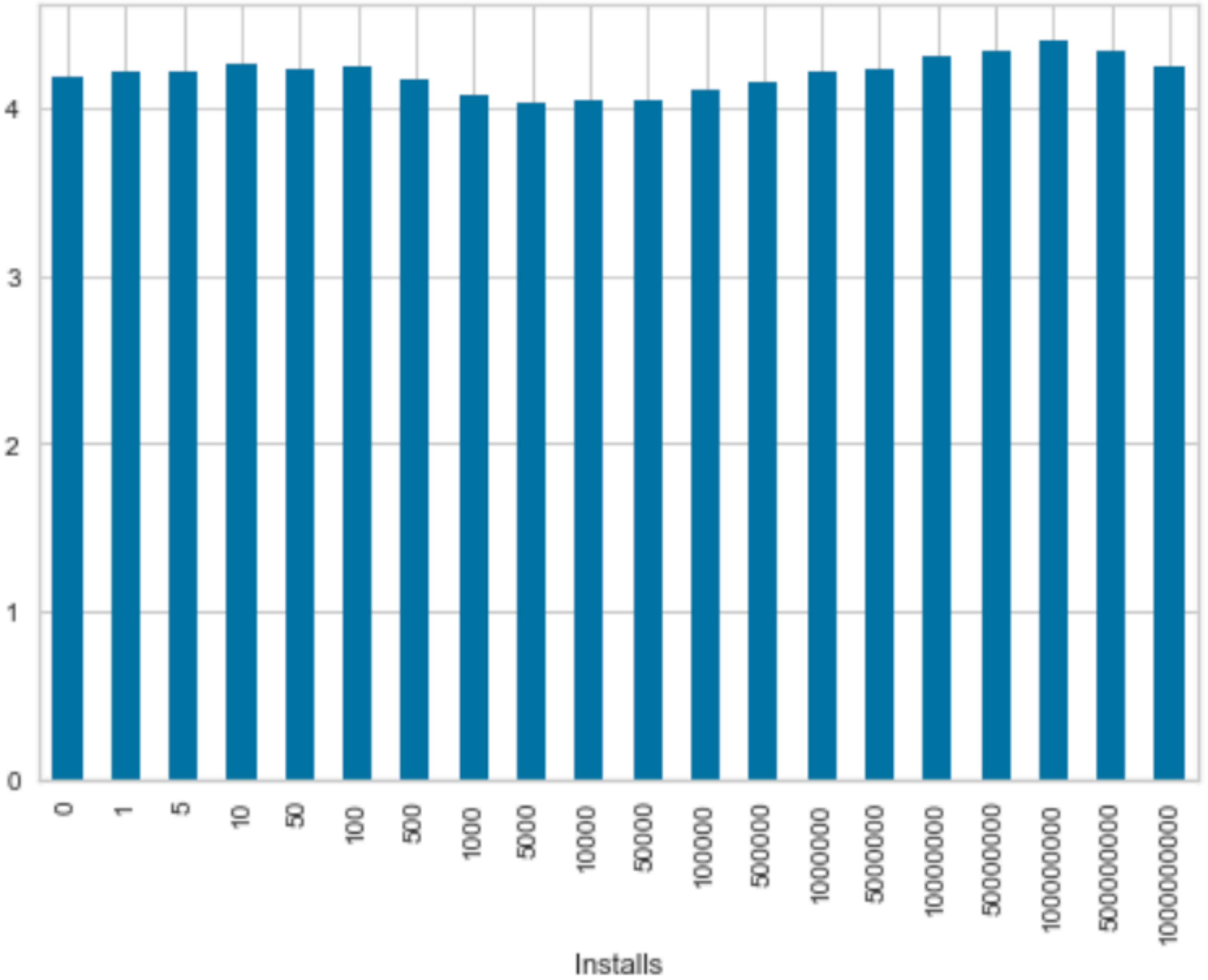


Distribution of app ratings, rounded. Ratings vary from 1-5.

Mean is 4.2



This graph dispalys the average rating for apps downloaded a certain number of times



We are now a bit familiar with the data.

Clusters: 5

Aggolomerative clustering

Kmeans clustering

Proceeded with KMeans

How apps were distributed over the clusters

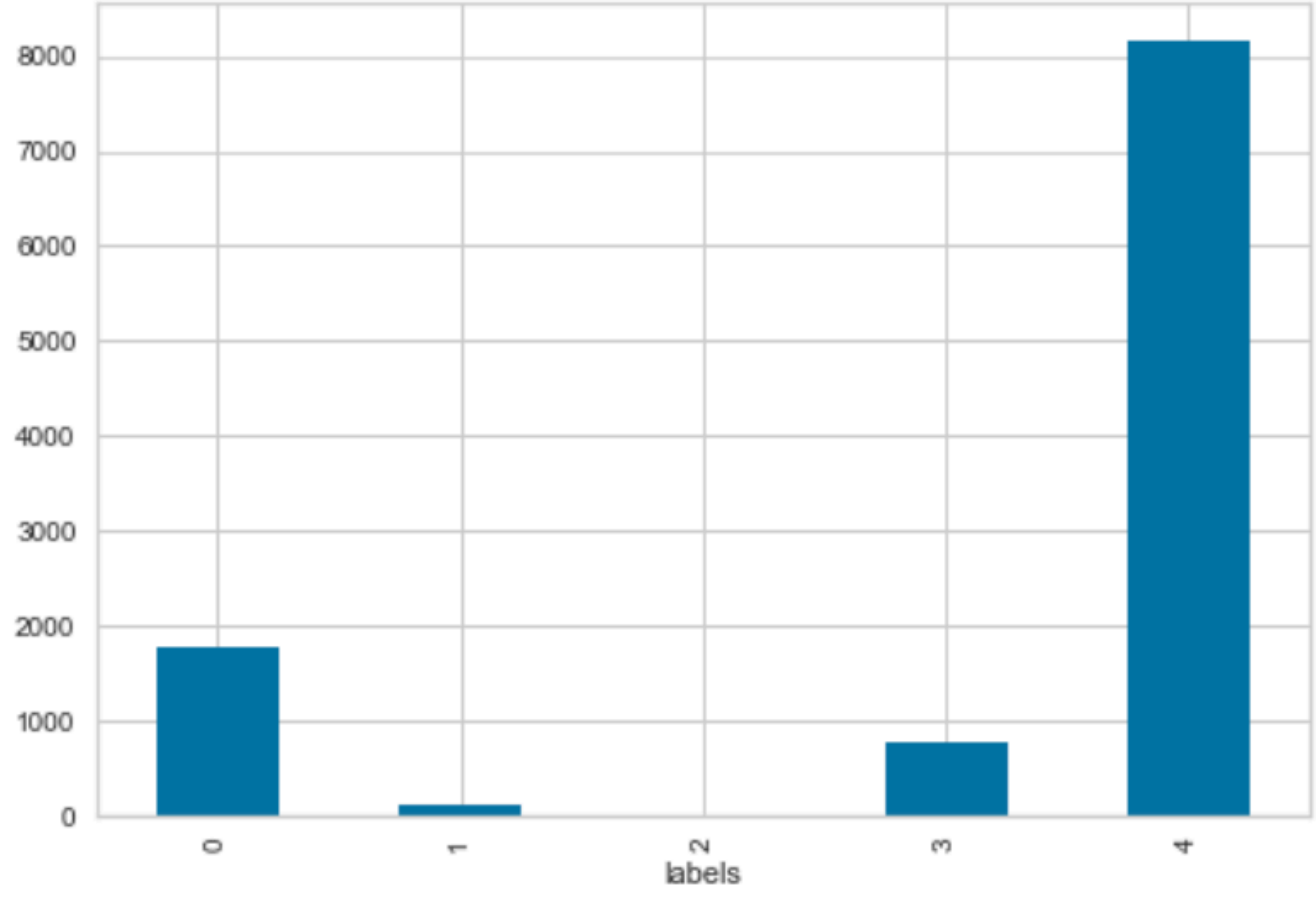
0: 1770

1: 115

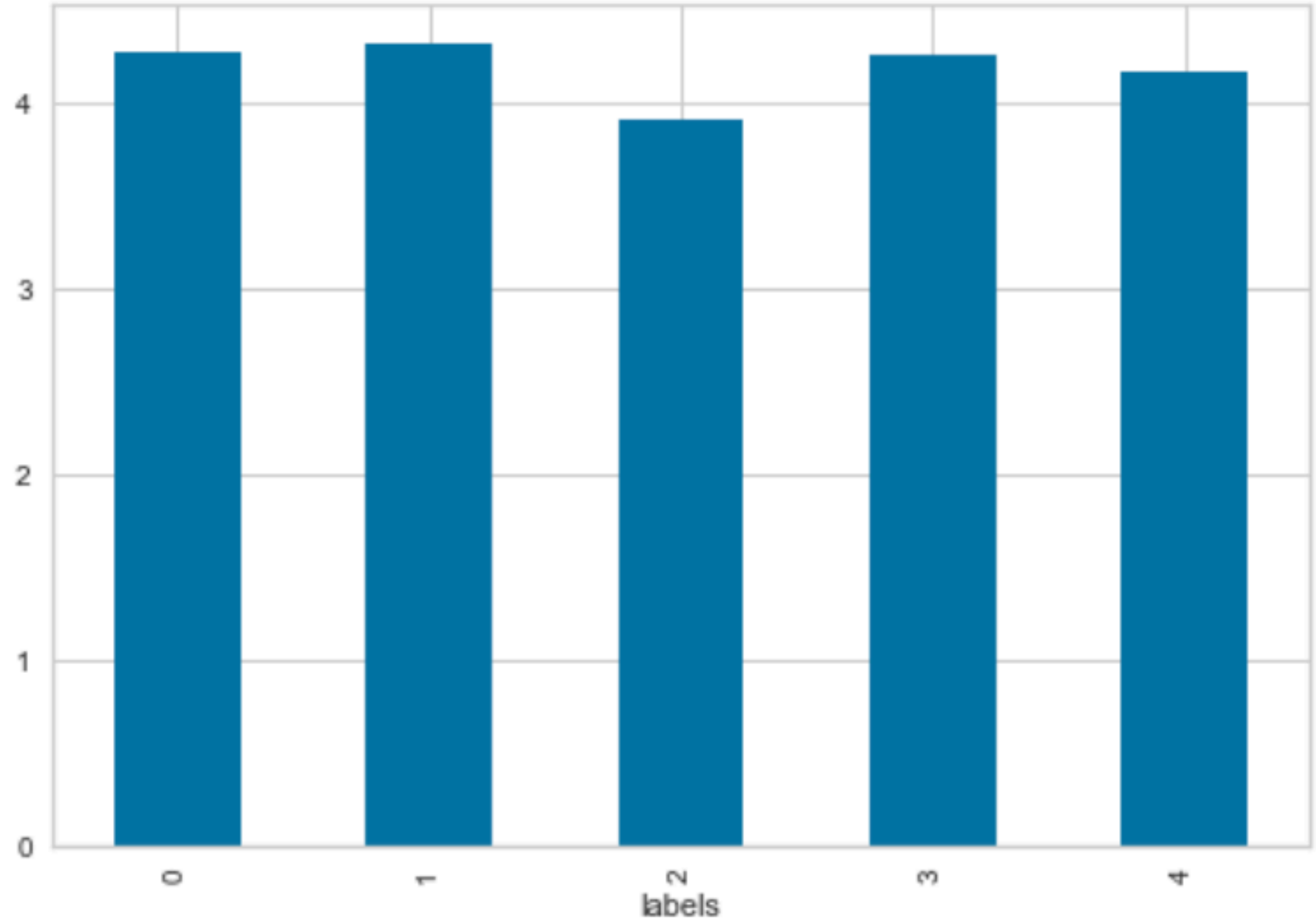
2:18

3: 779

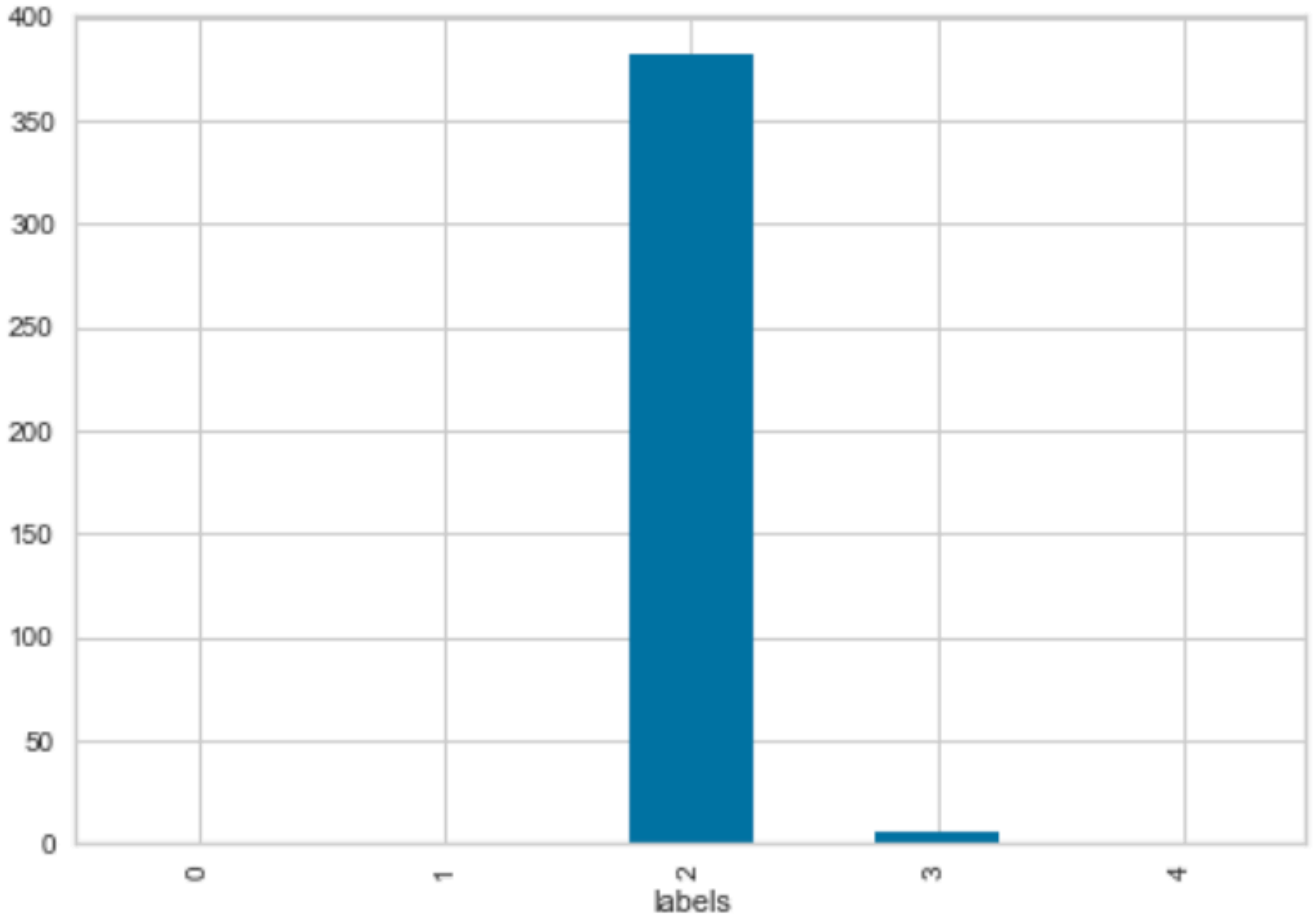
4: 8147



Mean rating of each cluster

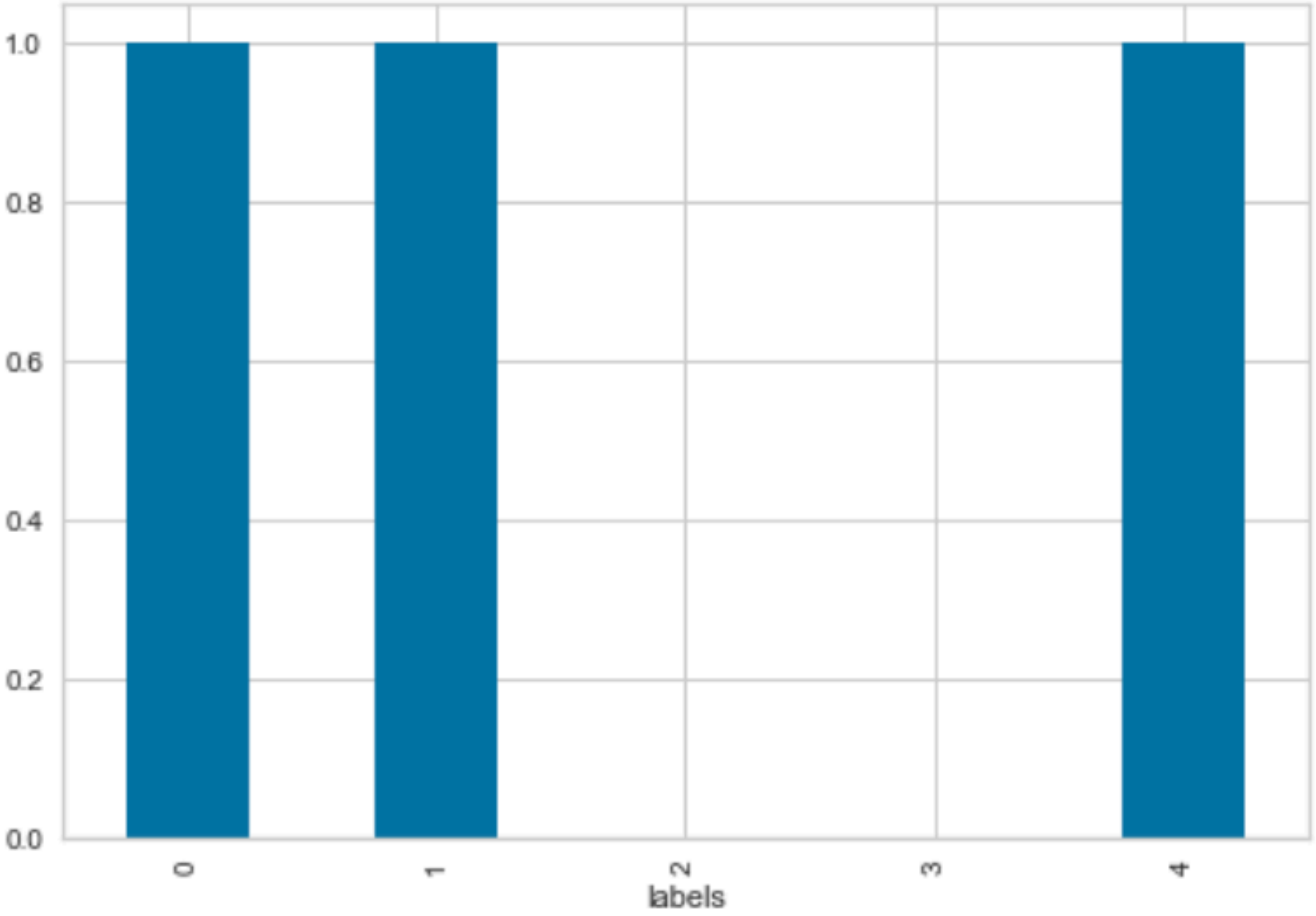


This graph displays average price of apps in the clusters.

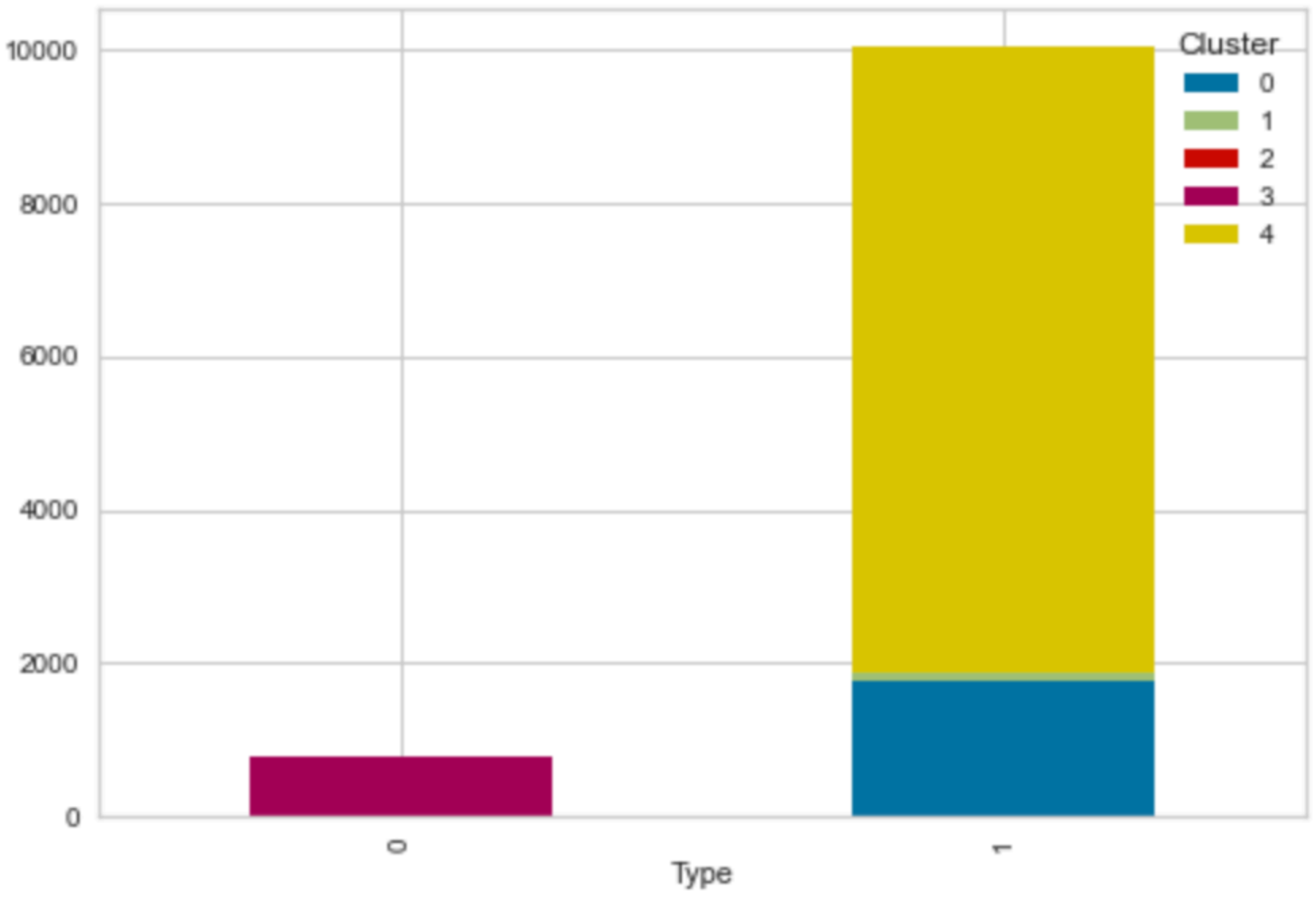


This graph displays the mean of 'Type' which is a column that gives an app 1 if it's free and 0 if you pay for it .

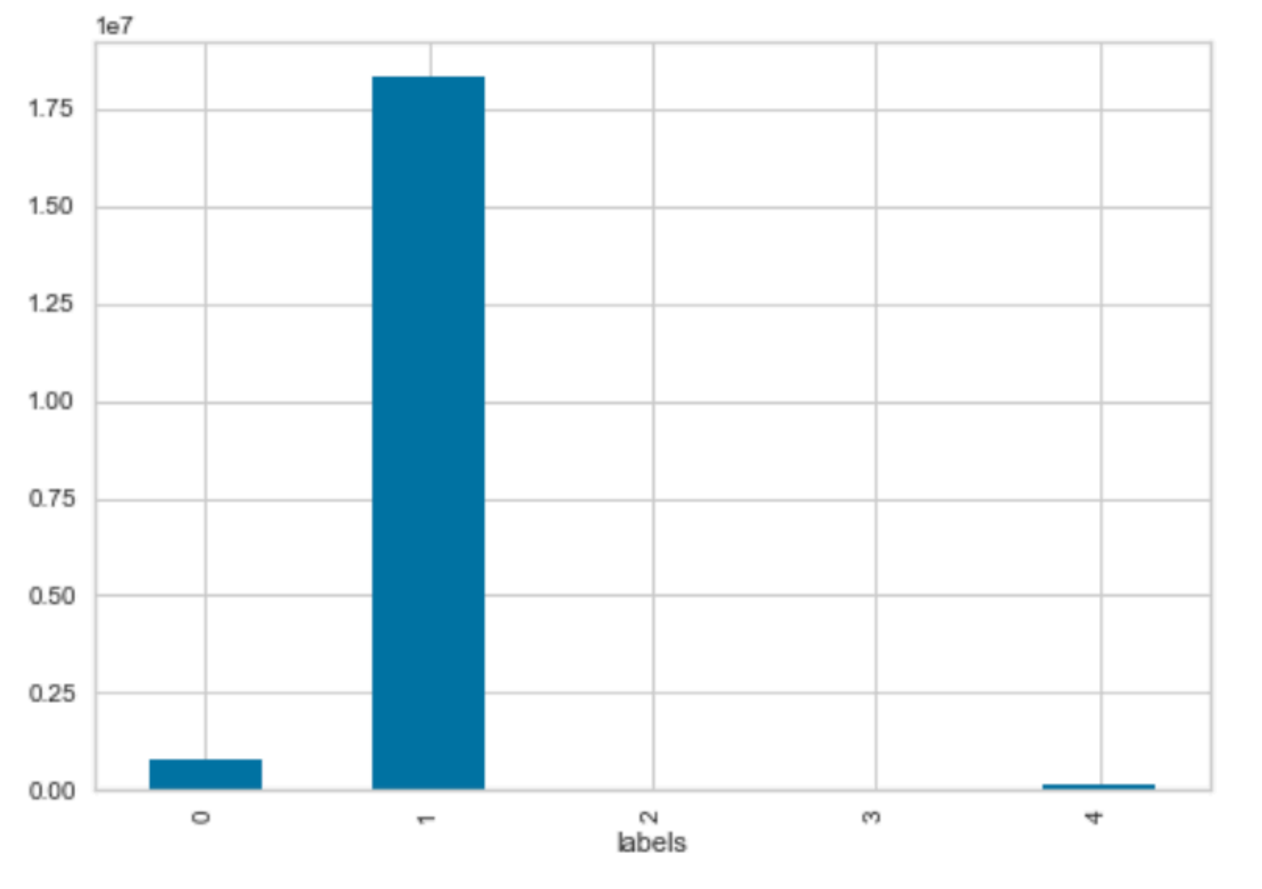
- Clusters 0,1 & 4 are all free apps
- Cluster 2, 3 are all paid apps



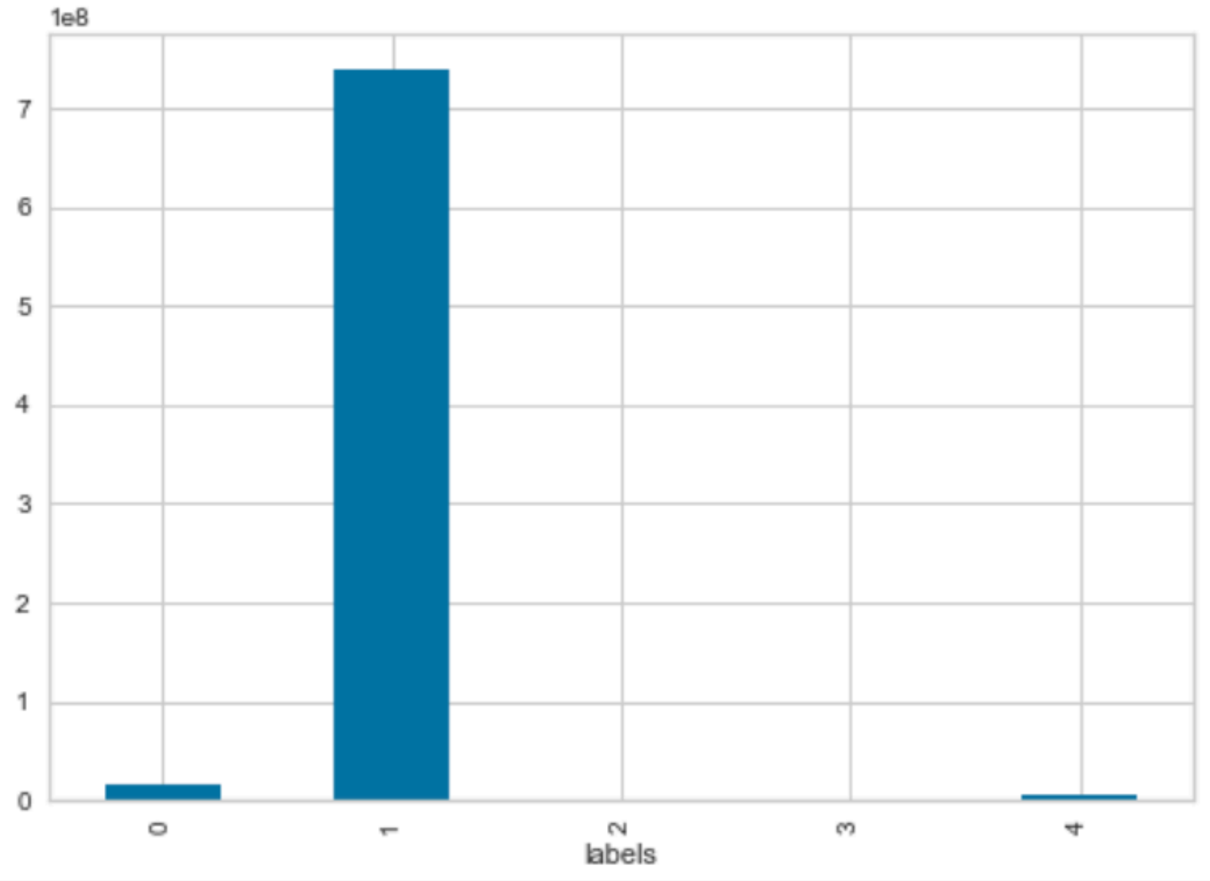
Again, free and paid apps are displayed. Left bar is paid apps and right bar is free apps.



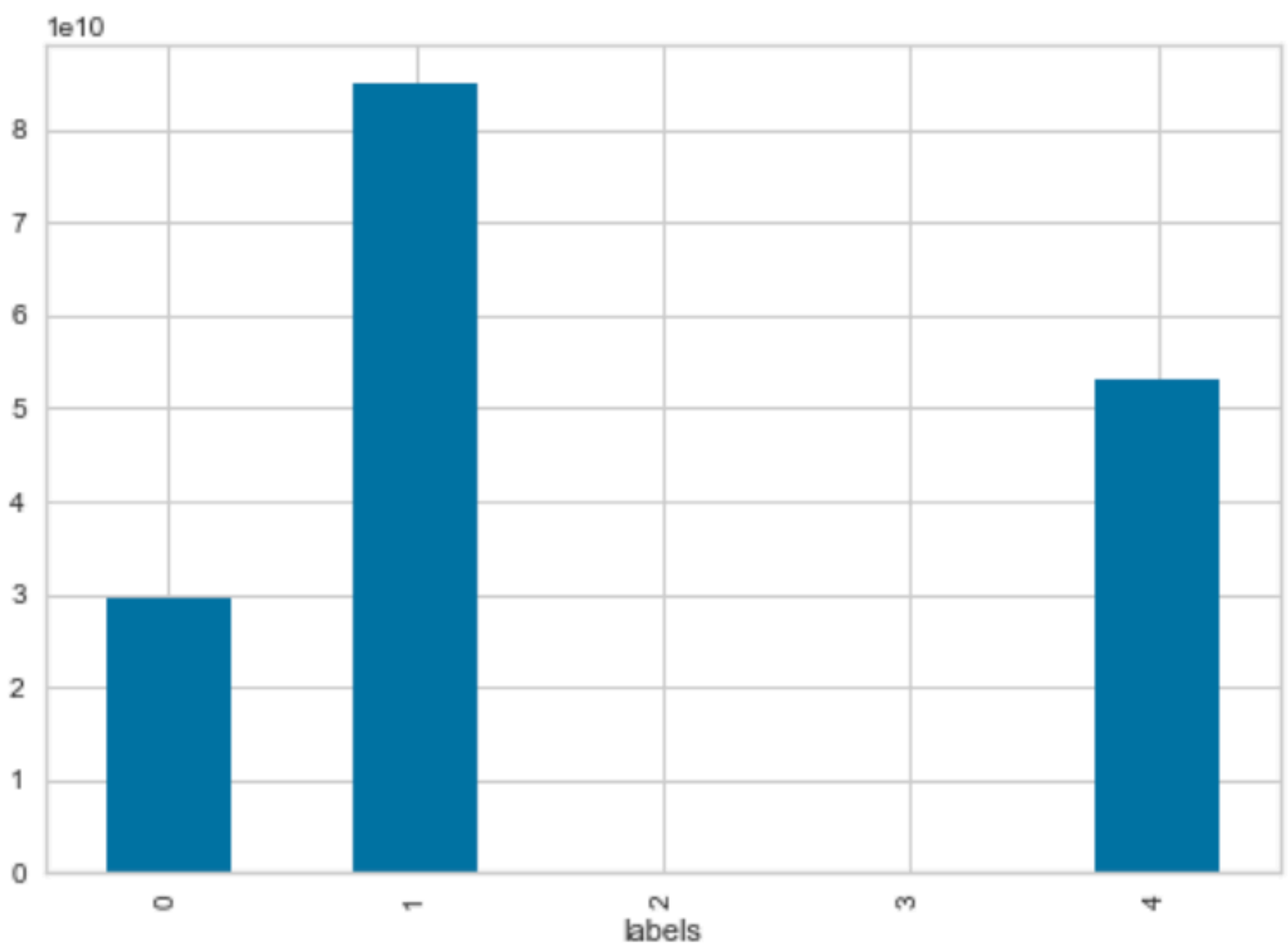
Average number of Reviews for apps in each cluster



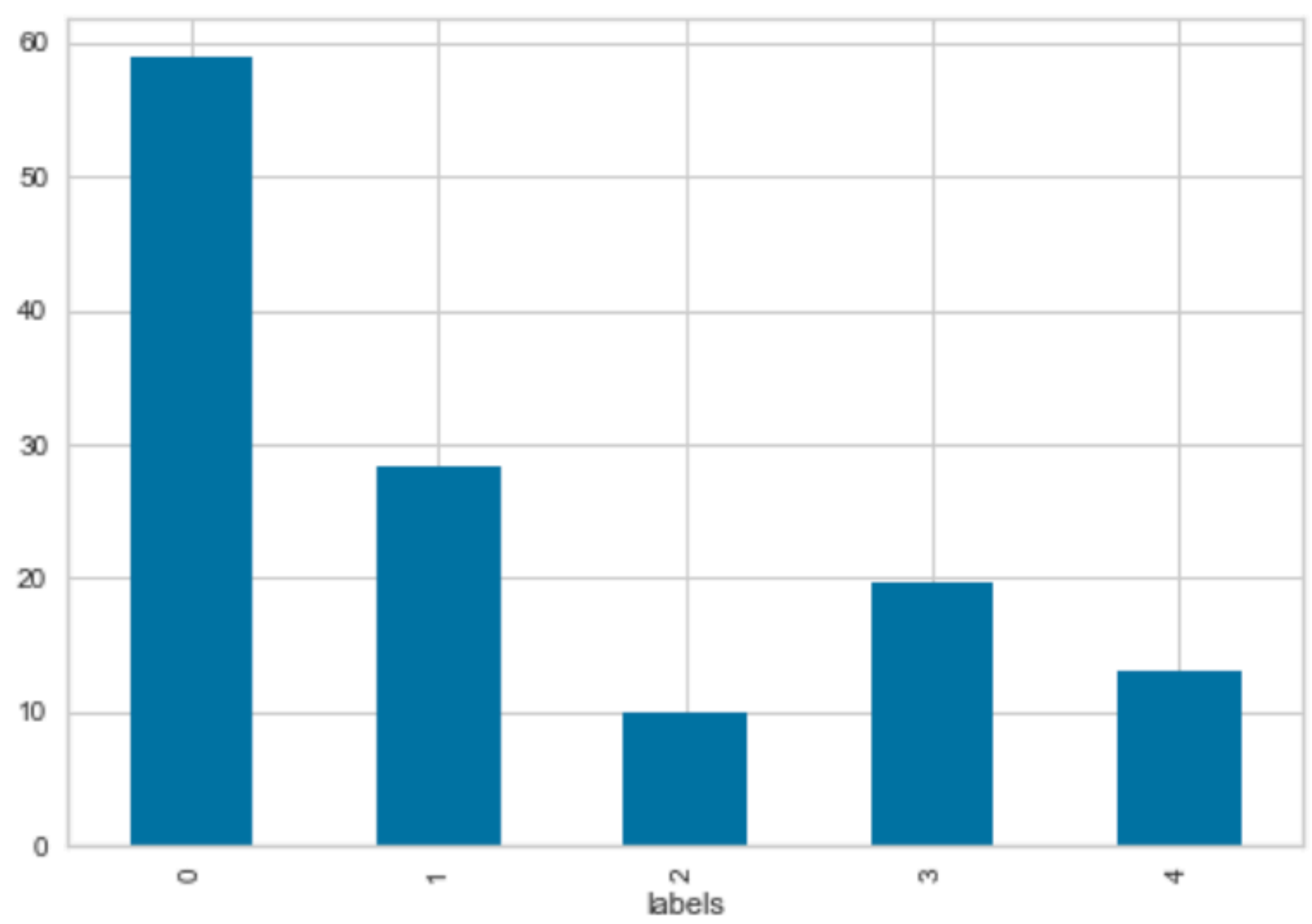
Average number of Installs for apps in each cluster



Total number of installs in each cluster. 2 & 3 are not nonexistent but so small in comparison to 0, 1, 4 they don’t show.

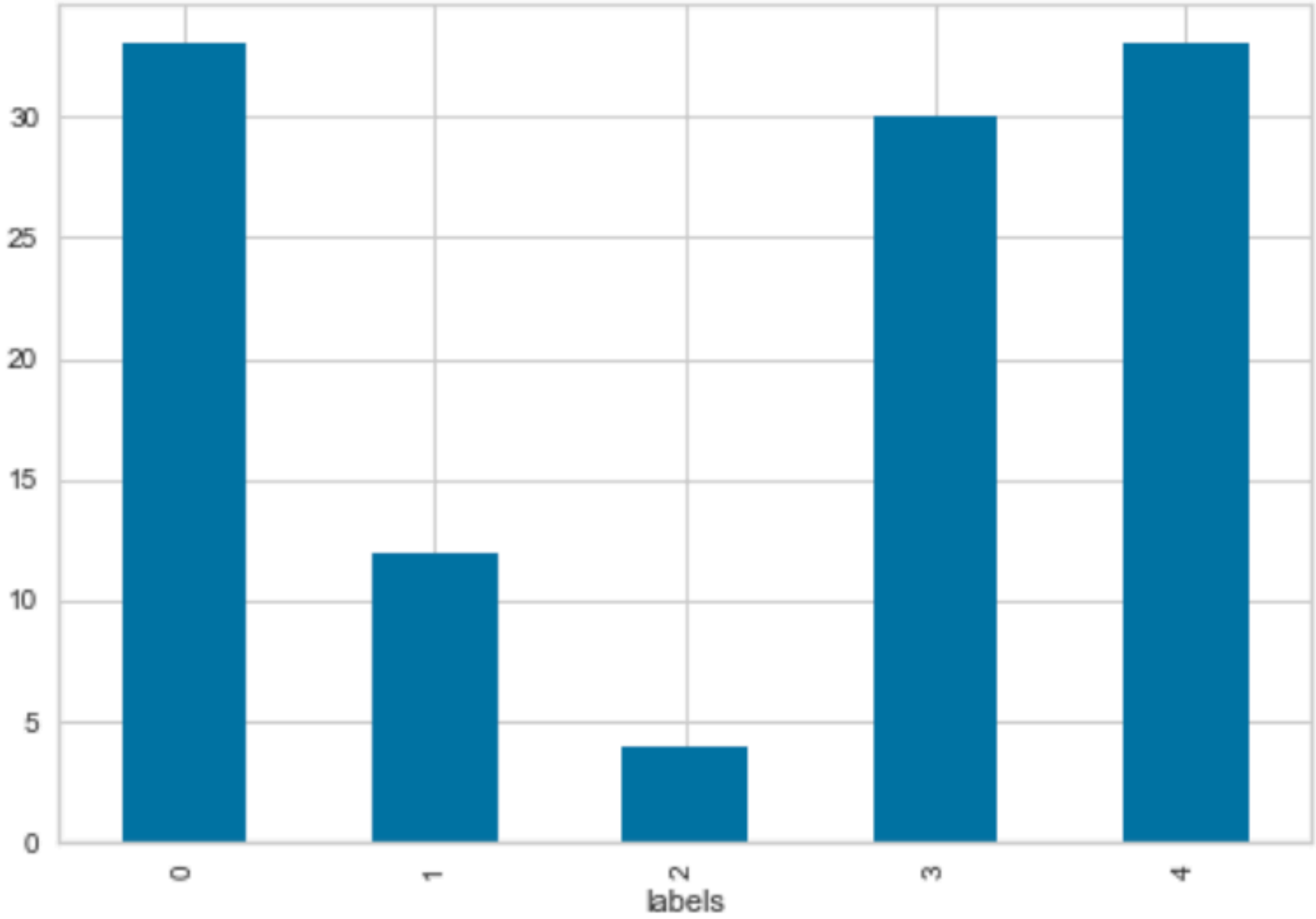


Average size of apps in each cluster in MB

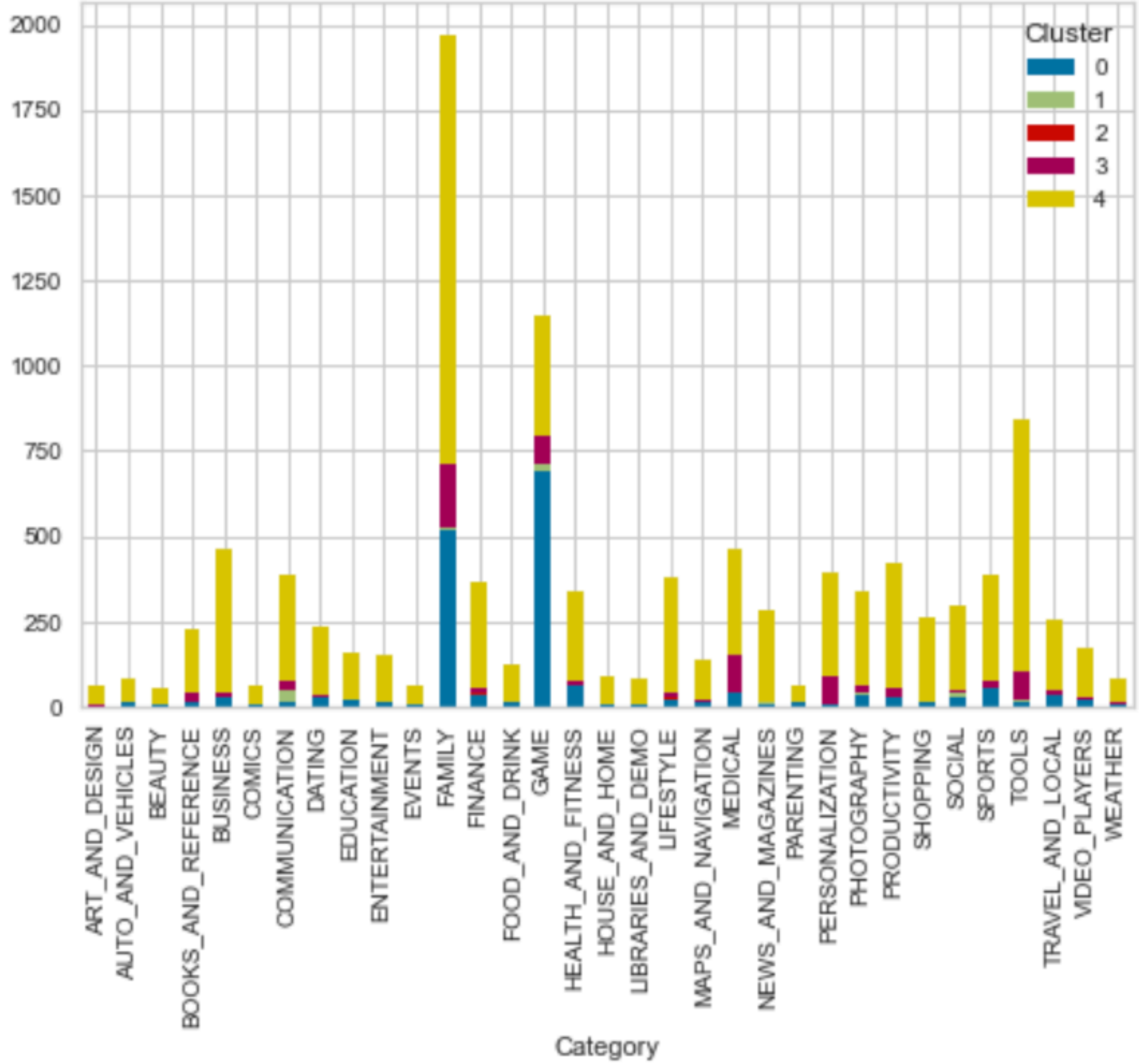


There are 33 clusters

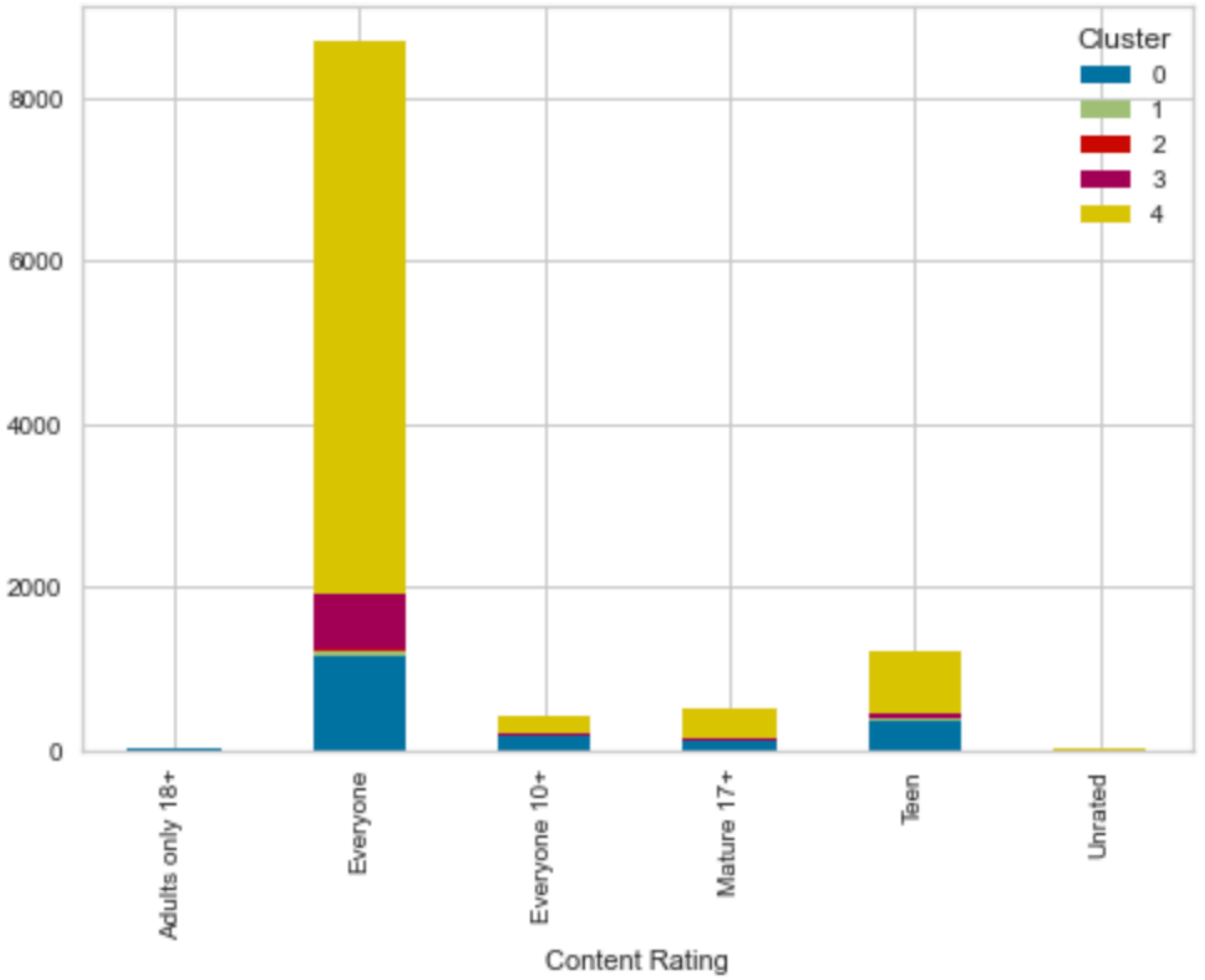
The graph displays how many categories are represented in each cluster



How apps in each category are distributed over the clusters



How apps in each Content Rating are distributed over the clusters.



Summary and cluster characteristics

0: 1770 apps, all free, cluster has 2nd most downloads and reviews, largest mean size of its apps and all categories are represented

1: 115 apps, all free, cluster has by far the largest avg installs and avg reviews but also total installs, 2nd largest mean app size and only 12 categories represented

2: 18 apps, all paid for (between 200-400$), tad bit lower average rating and only 4 categories represented

3: 770 apps, all paid for (under 200$), very few total installs, almost all categories represented (30/33)

4: 8147 apps, all free, all cats represented

THANK YOU!