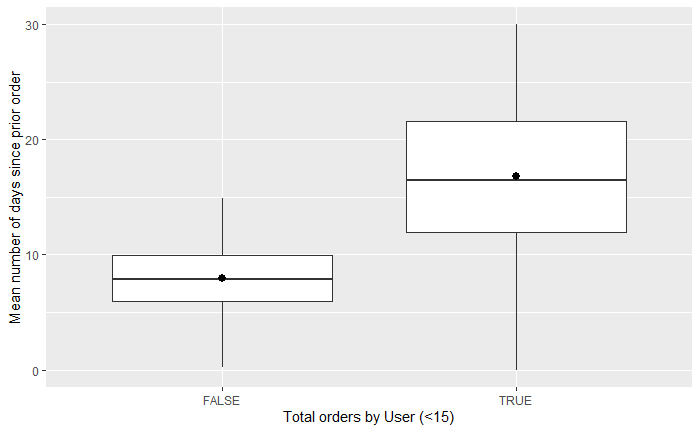
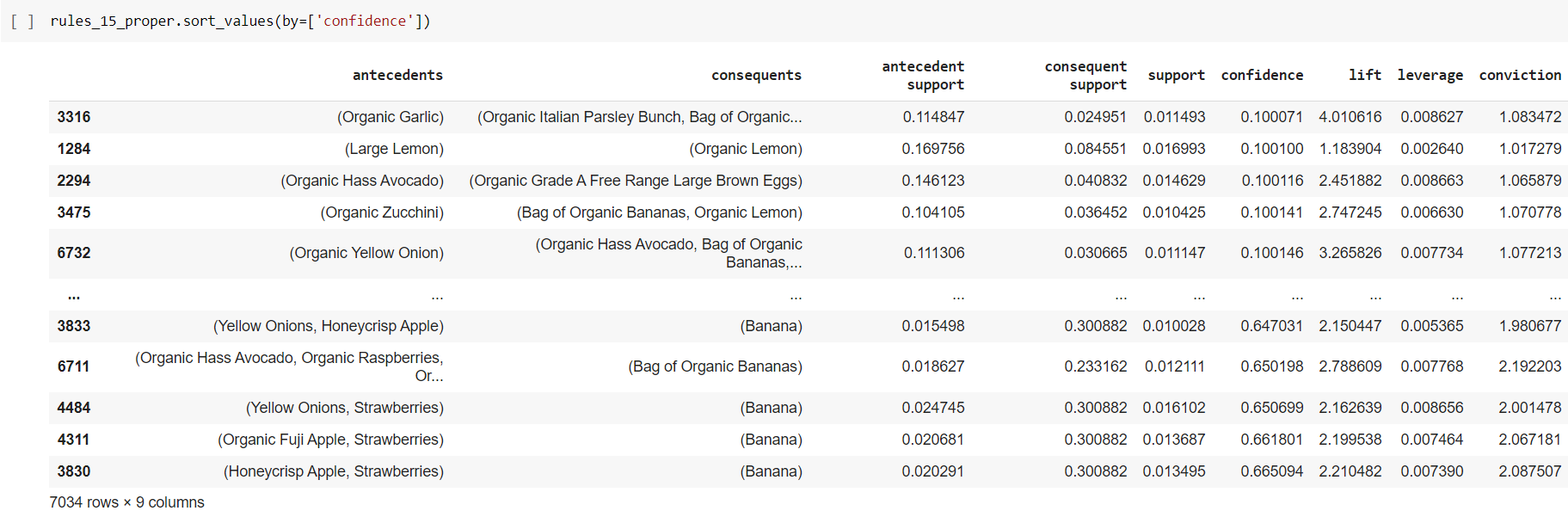
**Results**

The main goal of this project is to increase the revenue of Instacart. To achieve this, we plan to increase the user base of Instacart by targeting a particular user segment. In order to specifically target users who do not order frequently on Instacart, users who had ordered less than 15 times from instacart were segmented. These users showed a pattern of buying items at a larger gap than other users in the dataset.



25% of users who had ordered less than 15 times used to order with a gap of minimum 22 days. These were the target users for further analysis as making them regular customers could significantly increase the revenue and sales of instacart and all further analysis was on this target group. On applying market basket analysis on this segment of customers, multiple rules were generated that were able to provide insights into the buying pattern of the user segment.



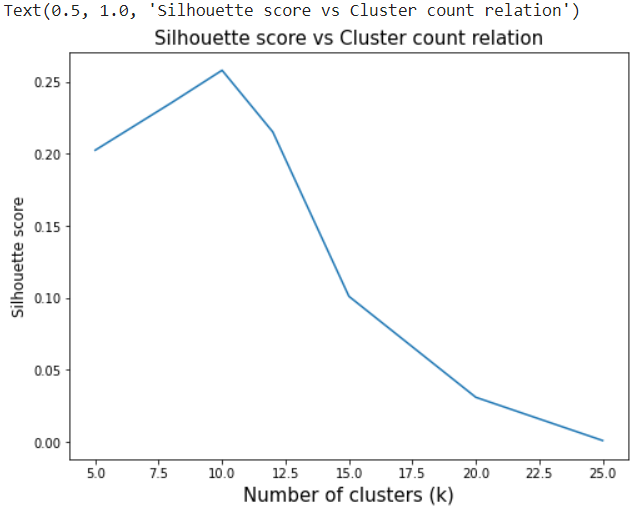
However, these rules still appeared to be influenced by a single common category of products, such as organic products or lime or bananas. So, this means, these rules were still generic and were not specific enough to be applied to the targeted user base. In order to obtain rules with more potential, it was decided that market basket analysis should be done on a more specific user base. It was concluded, further breakdown of users is necessary to find more targeted results that might help in increasing the revenue of the company.

To further break down the user segment, two approaches can be taken, one is K-means and another one is LDA. Here both methods are used to obtain user clusters. After obtaining the clusters, an average-sized user cluster is selected, and Market basket analysis is applied to that cluster using the apriori algorithm. Then rules obtained from both the clusters are compared to evaluate both the clustering techniques.

Before applying the clustering techniques, the total number of clusters required for K-means and the total number of topics required for LDA are determined using respective scores.

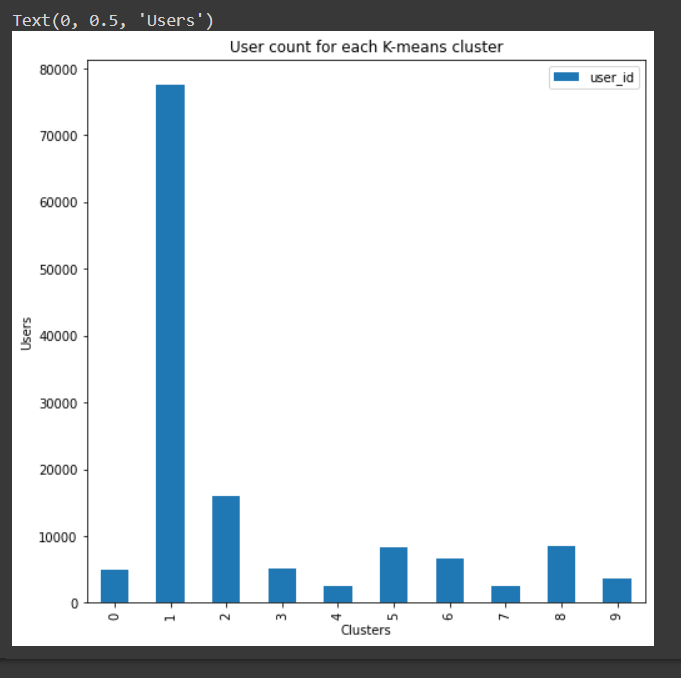
So, for K-means the total number of clusters (K) is determined by looking at the silhouette score. Below is the graph for all the different silhouette scores obtained for different values of K.

Silhouette score graph (K-Means Cluster count)

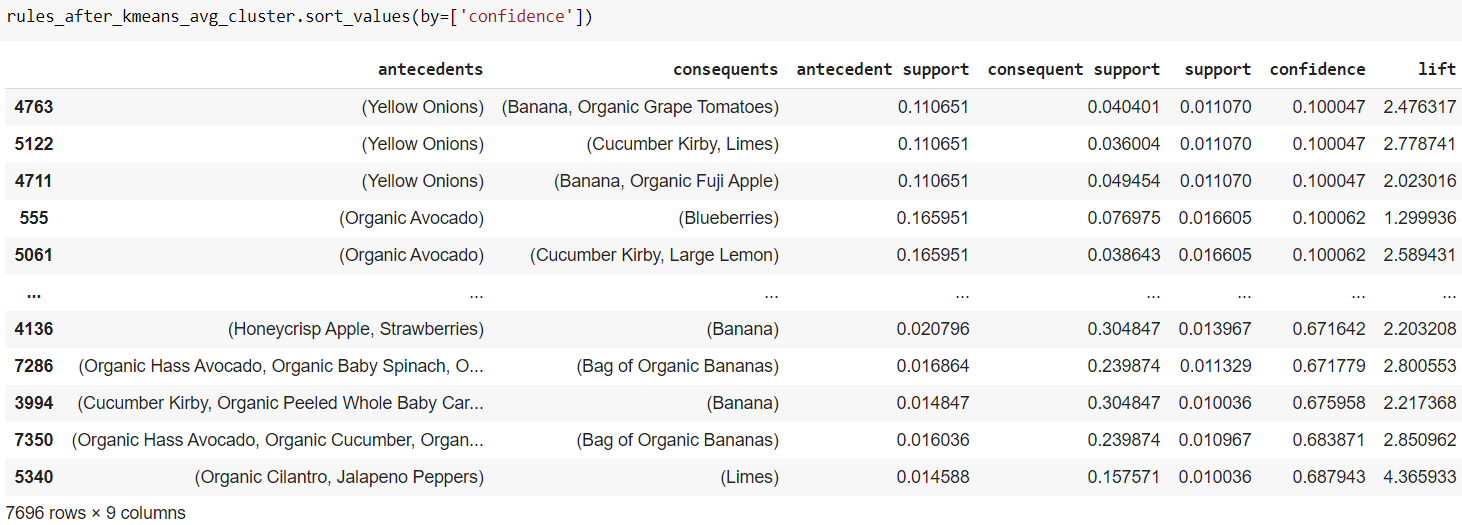


So according to the above graph, the value of K was decided to be 10. So, 10 clusters were obtained by applying K-means to the user segment. The count of users per cluster can be found out from the below bar plot.

User count for each K-means cluster



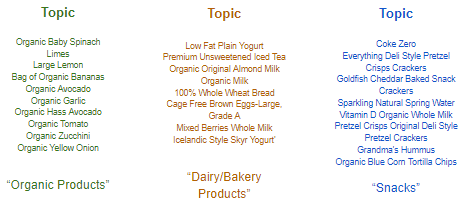
It was decided earlier that a cluster with an average user count was to be selected for the experiment. Hence, cluster number 2 was selected. After this, market basket analysis was applied to the K-means user clusters and the below rules were obtained.



However, these rules still appeared to be influenced by a single common category of products, such as organic products or lime or bananas. So, this means, these rules were still generic and were not specific enough to be applied to the targeted user base. Hence to improve the rules further, a probabilistic approach was taken and LDA was applied to the targeted user segment.

A more appropriate method to find groups of users with similar buying behaviour was to use LDA on the users where the set of documents would be replaced with a set of users and the number of words in each document would be replaced by the number of products bought by the user in their transaction records.

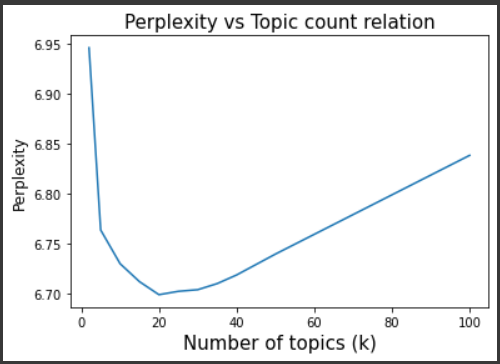
On the basis of the results of LDA, products which seem to share properties are in the same topic and hence these topics can be categorised on the basis of the intuitions of the users and many discernable patterns can be seen in the topics that have been modelled on the basis of the products in them. For example, the first topic clearly shows how the products with the highest probabilities are related to “Organic Products”, the second topic with “Dairy and Bakery Products”, et cetera.



Example of words with highest probabilities in topics at k = 20

For LDA the total number of topics (K) is determined by looking at the perplexity. Below is the graph for all the different perplexity values obtained for different values of K.

Perplexity graph (LDA topic count)

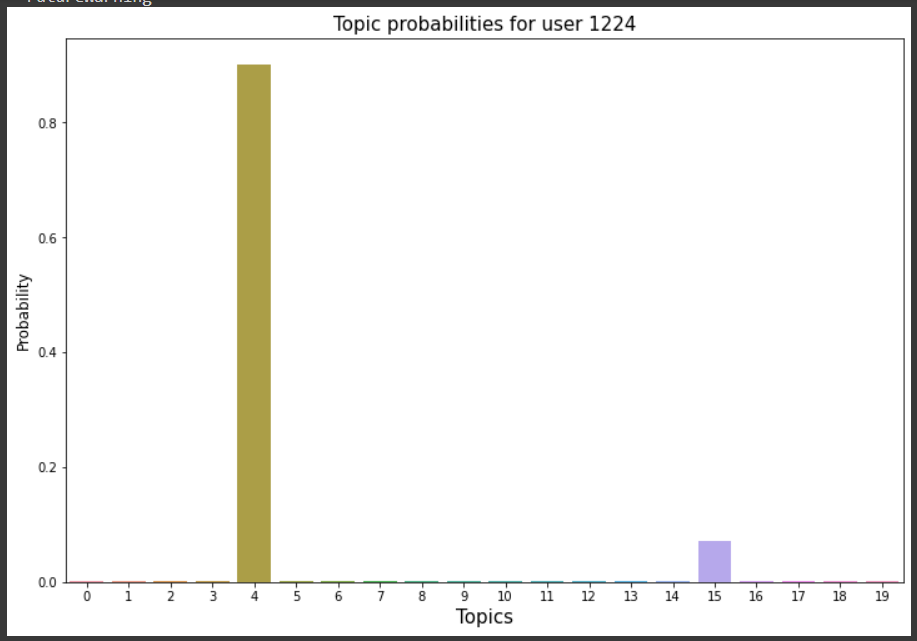
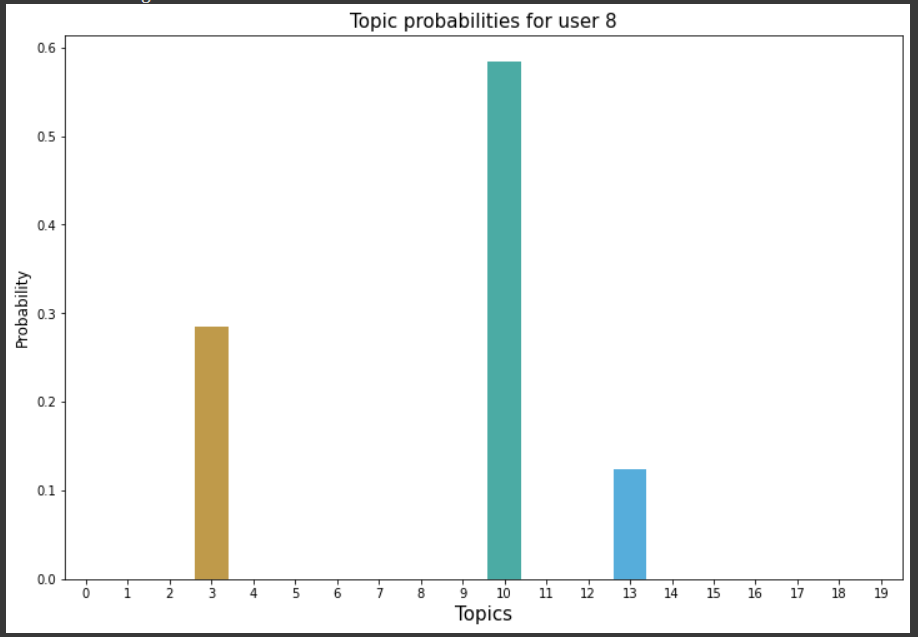


So according to the above graph, the value of K was decided to be 20. So, 20 topics were obtained by applying LDA to the user segment. Then clusters were obtained using the topic distributions for users.

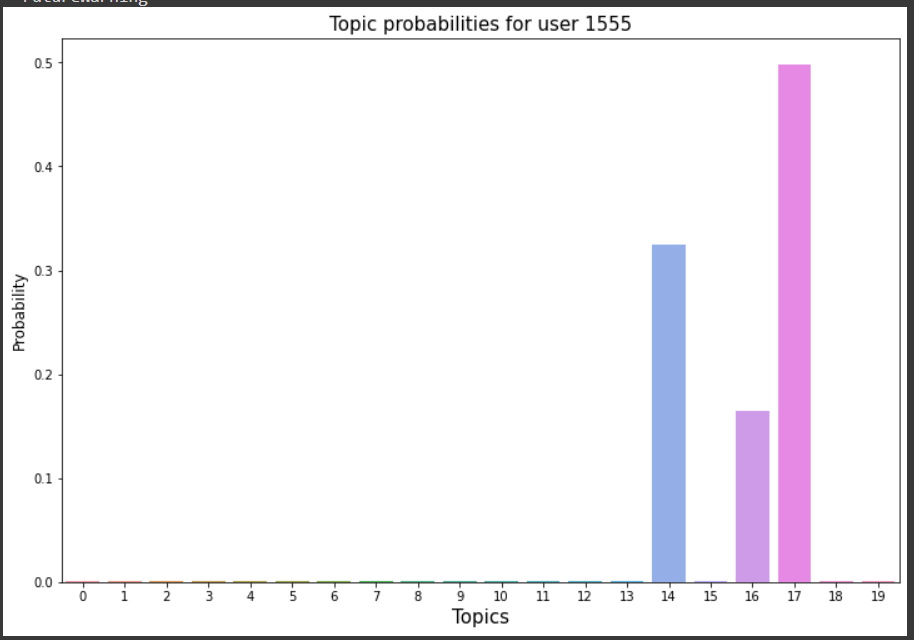
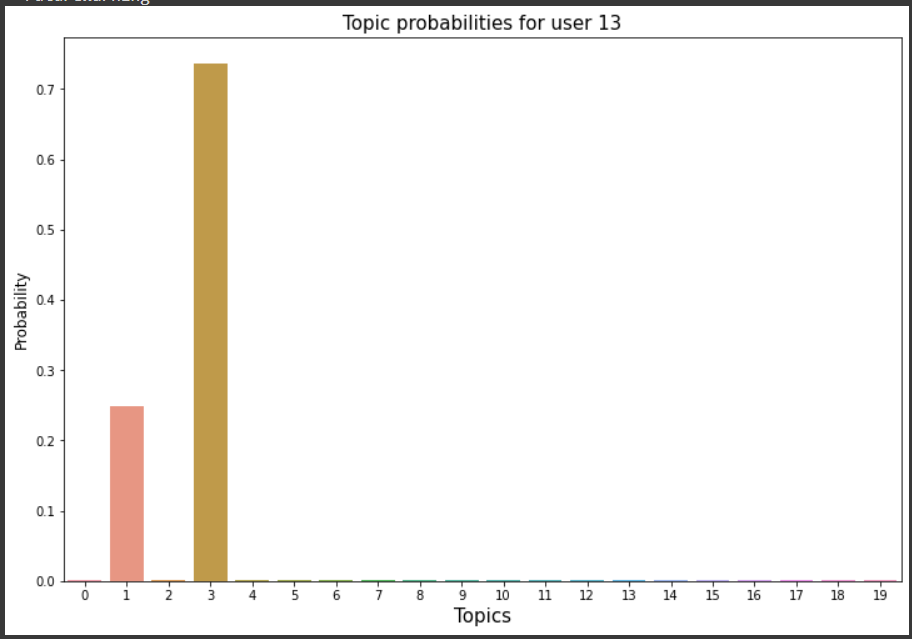
Just for a demo, topic distribution for 10 random users is displayed below. As LDA is applied, Dirichlet distribution is seen to some extent.

**User topic distribution**

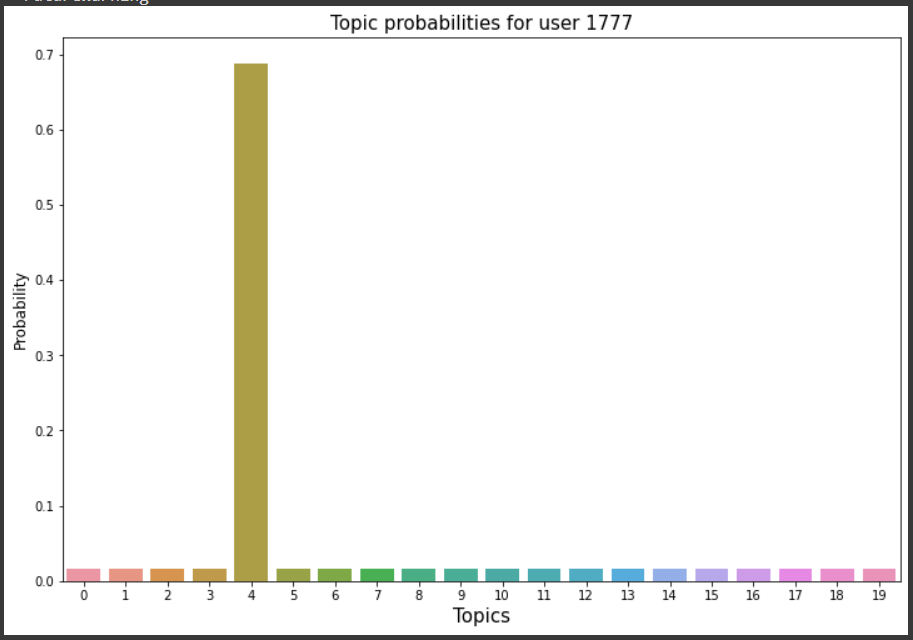
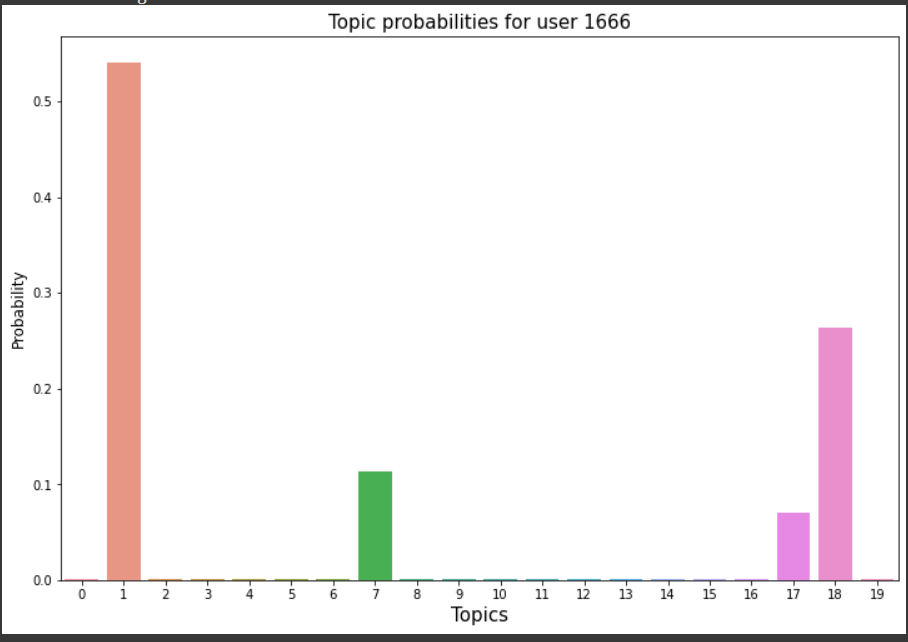
For user 8: For user 1224:



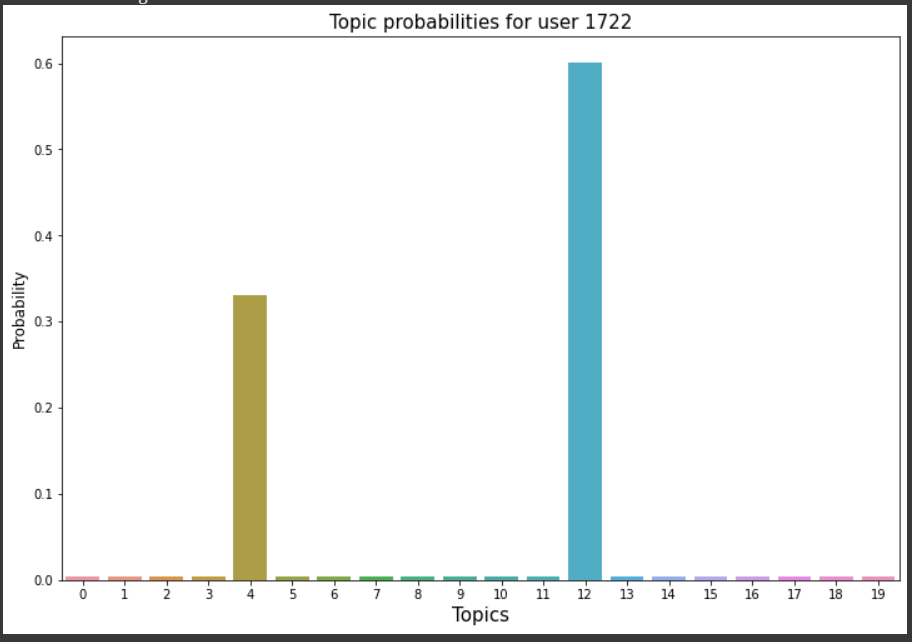
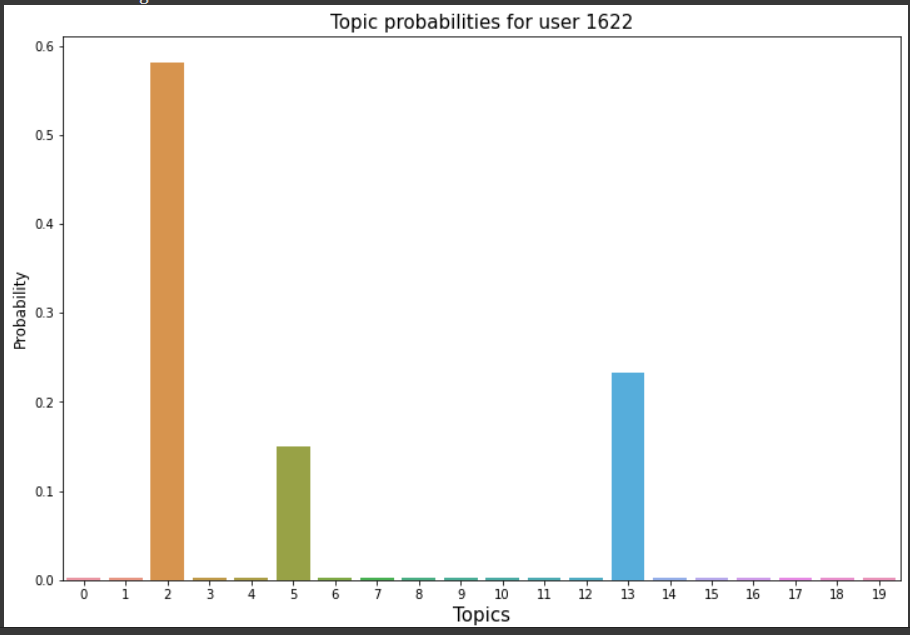
For user 13: For user 1555:



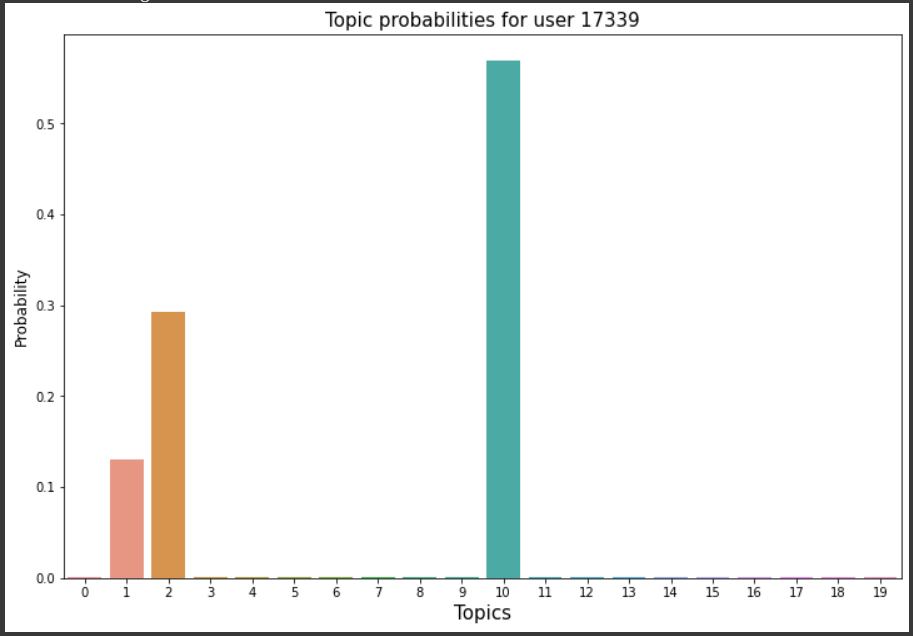
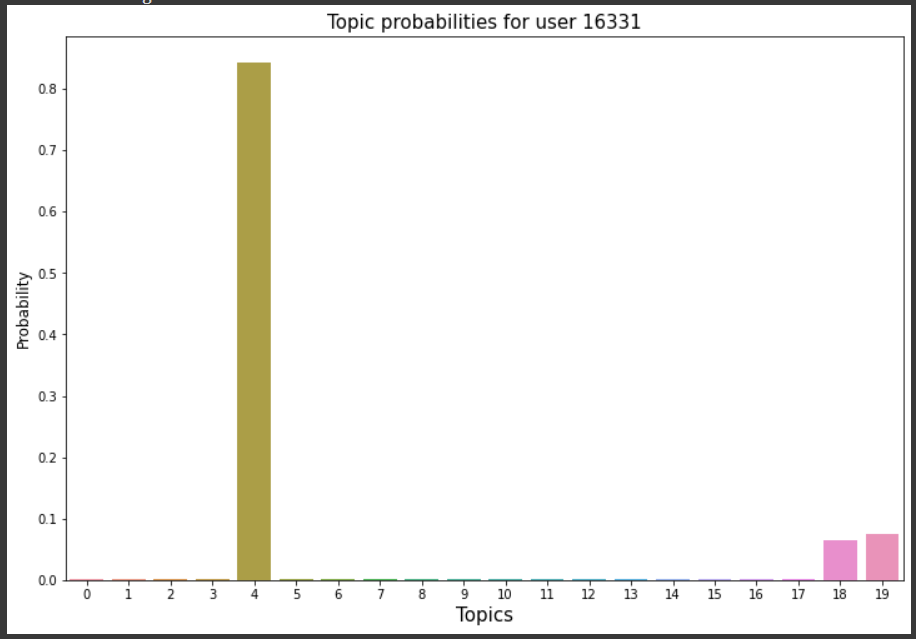
For user 1666: For user 1777:



For user 1622: For user 1722:

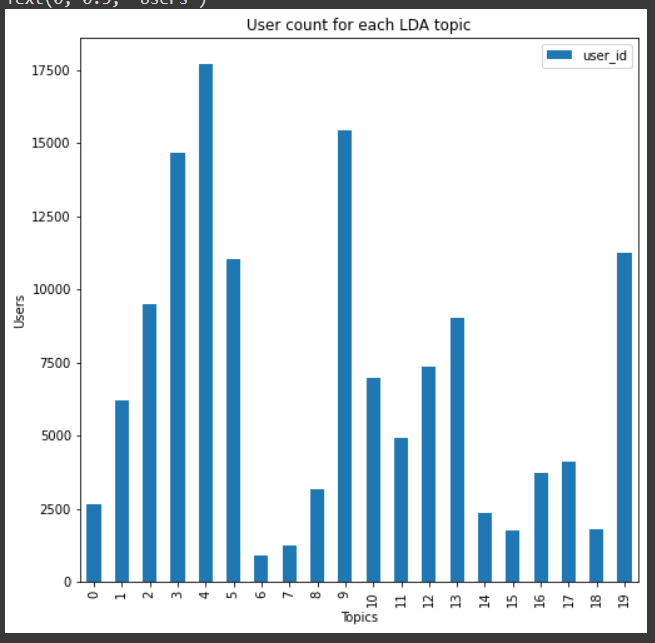


For user 16331: For user 17339:

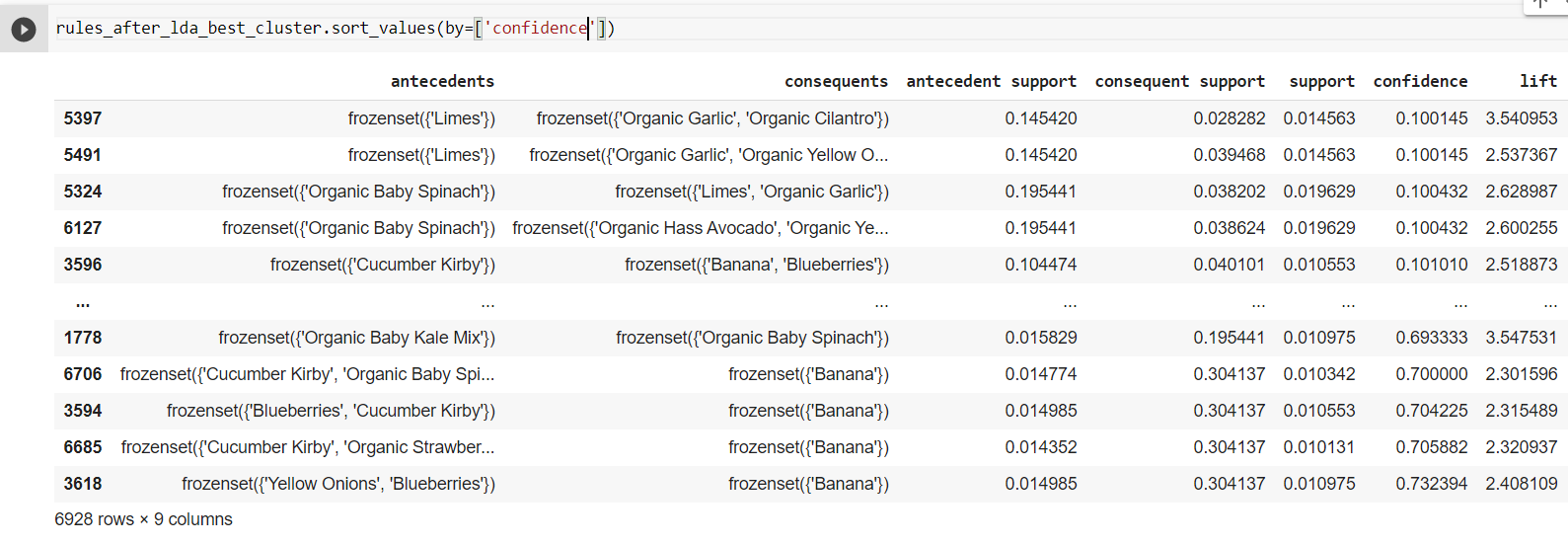


The count of users per topic can be found out from the below bar plot.

User count for each LDA topic



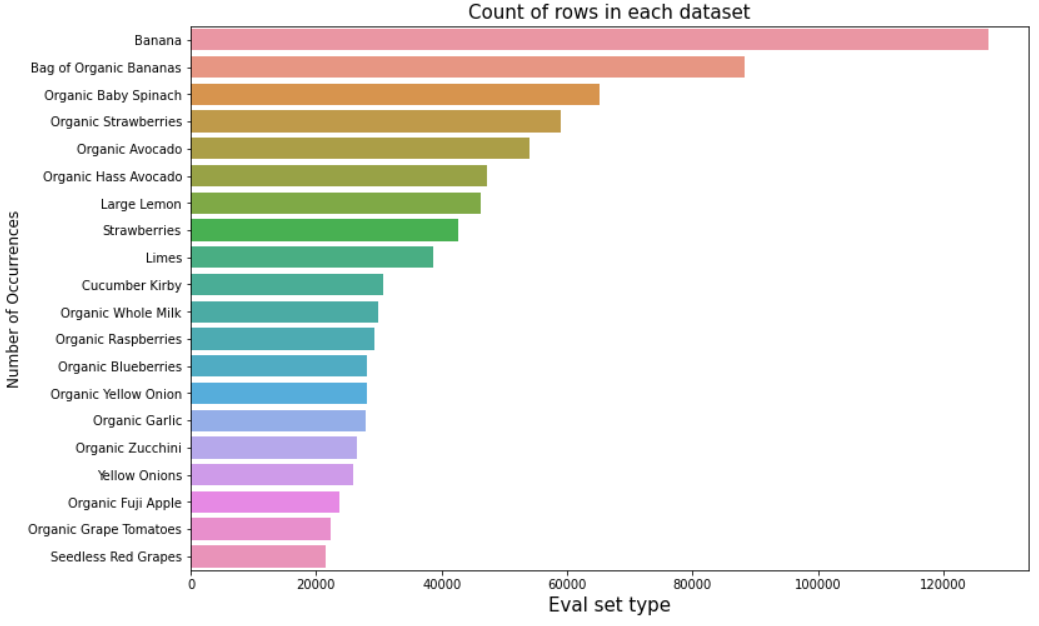
It was decided earlier that a topic cluster with an average user count was to be selected for the experiment. Hence, topic cluster number 14 was selected. After this, market basket analysis was applied to the LDA user topic clusters and the below rules were obtained.



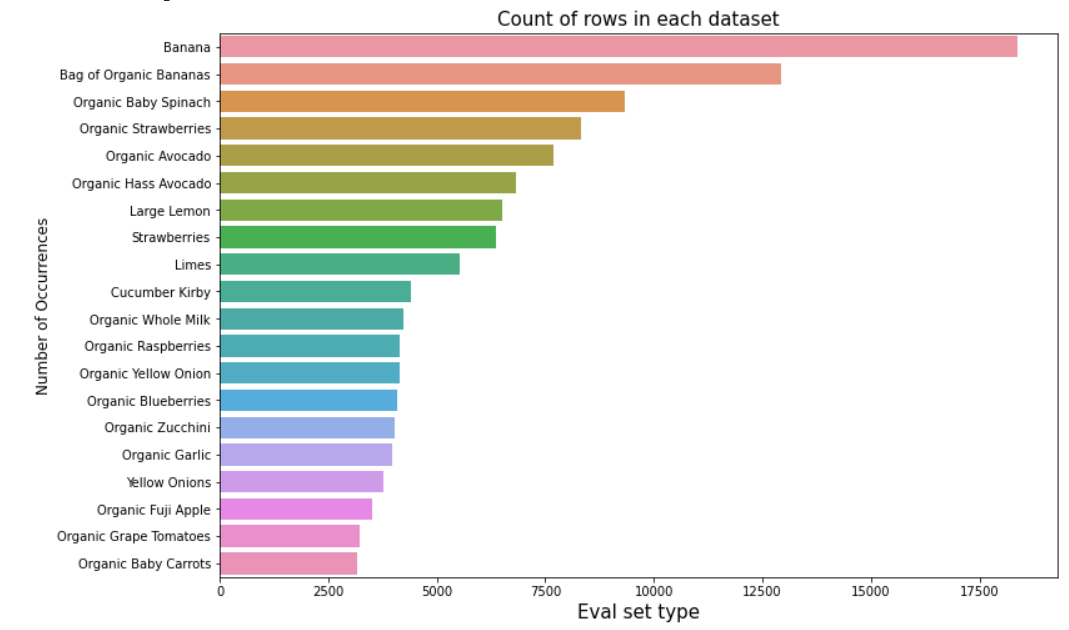
Finally, an improvement in the quality of rules can be seen after applying market basket analysis to the user cluster obtained as a result of LDA. These rules are not heavily influenced by a single common category of products, such as organic products or lime or bananas.

Hence it can be concluded from the results that taking a probabilistic approach can lead to obtaining rules with more potential. These rules can be used to create marketing models which can lure targeted users to order more products. Hence consequently increasing the revenue of the Company.

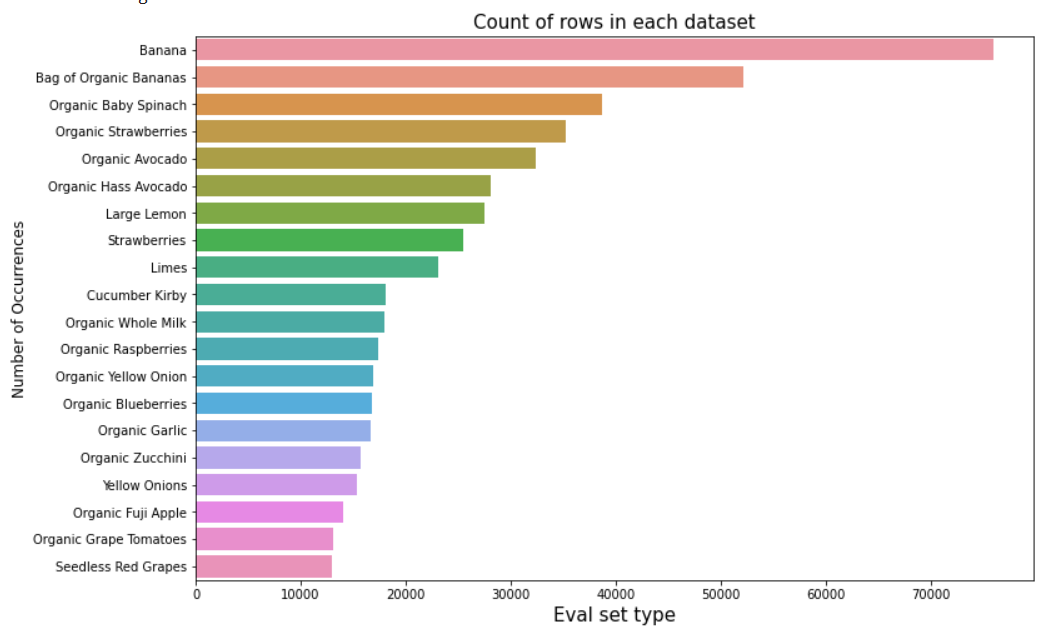
Less than 15 orders user cluster (User count = 136320)



K-means avg user count cluster (User count = 19332)



K-means biggest user count cluster (User count= 81055)



LDA avg user count cluster (User count = 4739) x(on 14)

