DATA SCIENCE AND BIG DATA ANALYTICS ASSIGNMENT

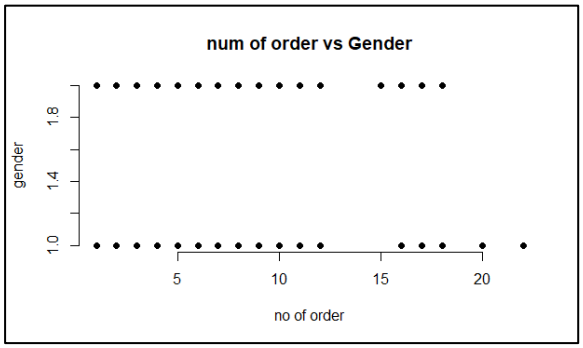
An Analysis Report On K-Means Clustering

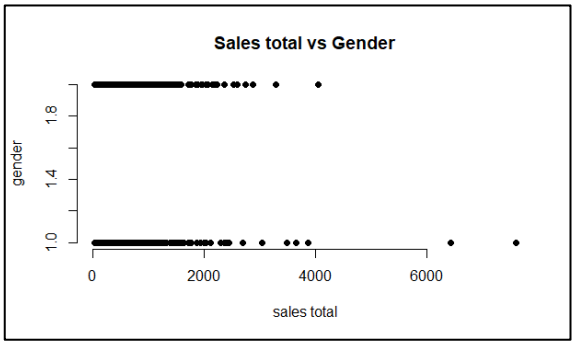
The given dataset comprises of details pertaining to sales with columns like customer id, sales total, no of order and gender pertaining to each and every customer who has bought a particular product or commodity. A total of 6269 entries are recorded in the dataset and the objective is to find appropriate group or cluster of customers based on the information given using K-means clustering algorithm.

The following is the head of dataset – includes first 6 values

	cust_id	sales_total	num_of_orders	gender
1	100001	800.64	3	F
2	100002	217.53	3	F
3	100003	74.58	2	M
4	100004	498.60	3	M
5	100005	723.11	4	F
6	100006	69.43	2	F

It is evident that cust_id is used to identify all the records uniquely and therefore doesn't influence other attribute. So it can be removed from further analysis. So we are now left with 3 other attributes for which we'll use scatter plot to determine the relationship between the variables or attributes.

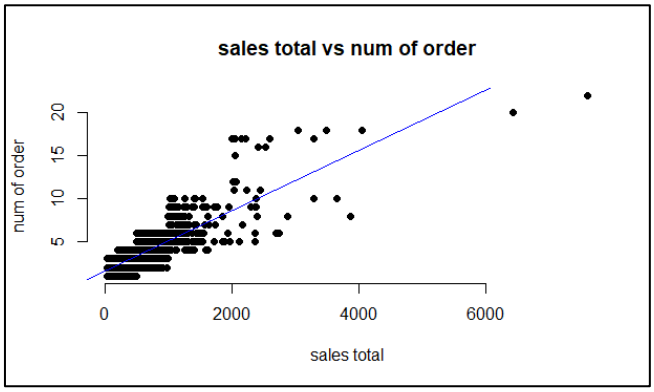




The graph depicts the correlation between the attribute gender with sales and order respectively and it is clear that there exists no correlation or weak correlation between the

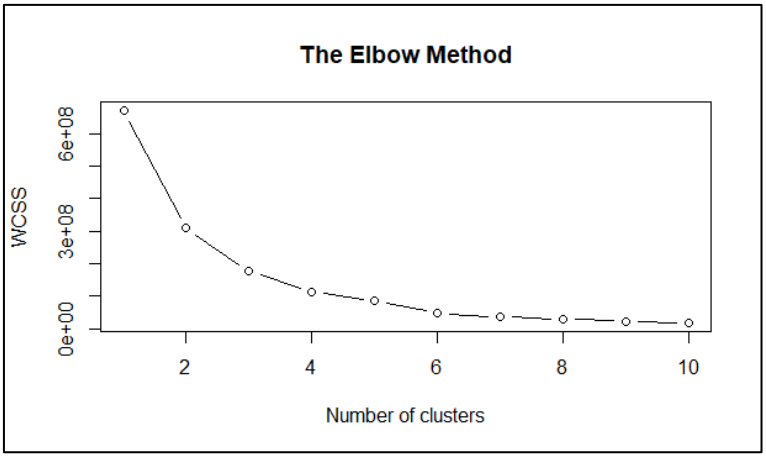
variables. Added to this gender is categorical (i.e. binary attribute) so can be removed since it doesn't offer much to explain the variance in data.

Now we are left with two attributes being sales total and no of orders whose distribution is as follows:



The graph shows a positive correlation which can be seen from the regression line and clearly these 2 attributes seem to be more correlated and we would apply the clustering technique on these 2 attributes.

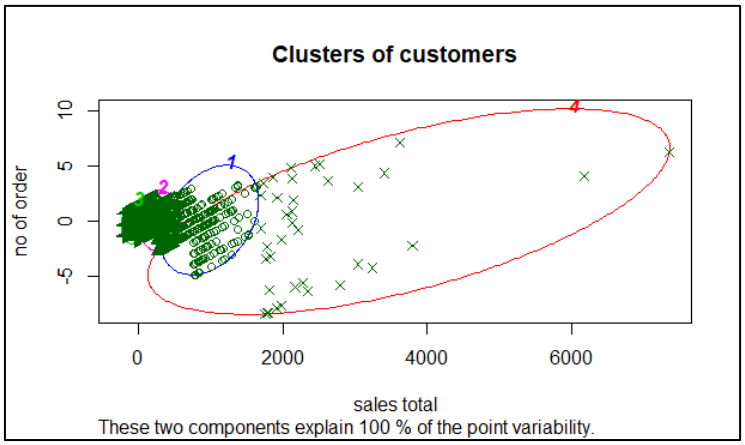
For K-Means to work efficiently it is important that we choose the K value appropriately and so we use the elbow technique to determine the suitable K value.



Within Cluster Sum of Square (WCSS) enables us to determine the appropriate value for K.

Here the line falls steeply till 4, after which it gradually decreases and hence we take K=4 as the ideal number of cluster for this algorithm.

After applying K-means with k=4 we can visualize the clusters-



The plot represents 4 clusters which are numbered with sales total and no of orders in x and y axis respectively. It is evident from the clustering that the majority of sales is obtained from clusters 2 and 3 which consists of less number of orders (btw 0 and 3) and the sales total being less as well with less than 1000. Targeting this group and providing discounts to other groups might enhance sales.

Done By :

S.Ananya

CSE-E

RNO-088