DATA MINING

Assignment-2.2

SOBEY'S SALES DATA - CLUSTERING

Fatima Parveen
Meenal Shah
Navjot Kaur

25th January, 2018

INTRODUCTION

This report is on Sobey's sales data - Clustering. As we want to cluster customers based on revenue, number of visits and number of items bought from Sobey's store. The same attributes (revenue, visits, items) are considered for Cluster Profiling.

DATA PREPARATION

Below is the table that contains columns Customer_ID, HighSpent, LowSpent, HighVisits, LowVisits, HighItems, LowItems.

Customer_ID	HighSpend	LowSpend	HighVisits	LowVisits	HighItems	LowVisits
59583005	53.18888	-53.18888	11	-2	152	-69
36172216	369.589	-50.18889	9	-1	152	-70
58232932	62.79666	-62.79666	9	-2	151	-61
61290435	117.4122	-117.4122	8	-2	149	-67
59564012	54.46888	-54.46888	7	-2	148	-59
58638337	53.59333	-53.59333	6	-1	147	-68
39755604	378.8921	-51.45221	12	-2	146	-67
21566188	409.2153	-55.56999	6	-1	146	-66
21602722	57.79999	-57.79999	7	-1	145	-66

FINDING OUTLIERS

The above table has outliers i,e high and low values that are affecting other observations. So we calculated following values for all attributes and found Z-Score. No values falling above Z-Score of 2.68 are considered and all values below Z-Score -2.68 are eliminated, since these values are above the third quartile or below the first quartile.

Calculating Mean, Standard Deviation, Mean, Maximum, Minimum, and Range to find Z-score:

	HSpent	LSpent	HVisits	LVisits	HItems	LItems
SD	282.0457	106.6202	5.77041	1.576333	31.73499	17.07421
Mean	281.7529	-138.737	6.837045	-2.58097	67.08806	-24.4342
Max	2084.478	-50.0511	46	2	247	-5
Min	50.05111	-683.82	0	-16	17	-125
Range	2034.427	633.7687	46	18	230	120

Customer_	HighSpent	Zscore	LowSpent	Zscore	HighVisits	Zscore	LowVisits	Zscore	HighItems	Zscore	LowIter	Zscore
36172216	369.5890138	0.31143	-50.18888778	0.831	9	0.375	-1	1.003	152	2.6757	-70	-2.66869
59583005	53.18888444	-0.81038	-53.18888444	0.802	11	0.721	-2	0.369	152	2.6757	-69	-2.61012
41538512	768.961028	1.72741	-310.0488189	-1.607	5	-0.32	-1	1.003	137	2.203	-69	-2.61012
58638337	53.59332556	-0.80895	-53.59332556	0.799	6	-0.15	-1	1.003	147	2.5181	-68	-2.55156
61290435	117.4122144	-0.58267	-117.4122144	0.2	8	0.202	-2	0.369	149	2.5811	-67	-2.49299
39755604	378.892053	0.34441	-51.45220778	0.819	12	0.895	-2	0.369	146	2.4866	-67	-2.49299
59504655	123.5833078	-0.56079	-123.5833078	0.142	7	0.028	0	1.637	131	2.0139	-67	-2.49299
21566188	409.2152736	0.45192	-55.56999444	0.78	6	-0.15	-1	1.003	146	2.4866	-66	-2.43442
21602722	57.79998667	-0.79403	-57.79998667	0.759	7	0.028	-1	1.003	145	2.4551	-66	-2.43442

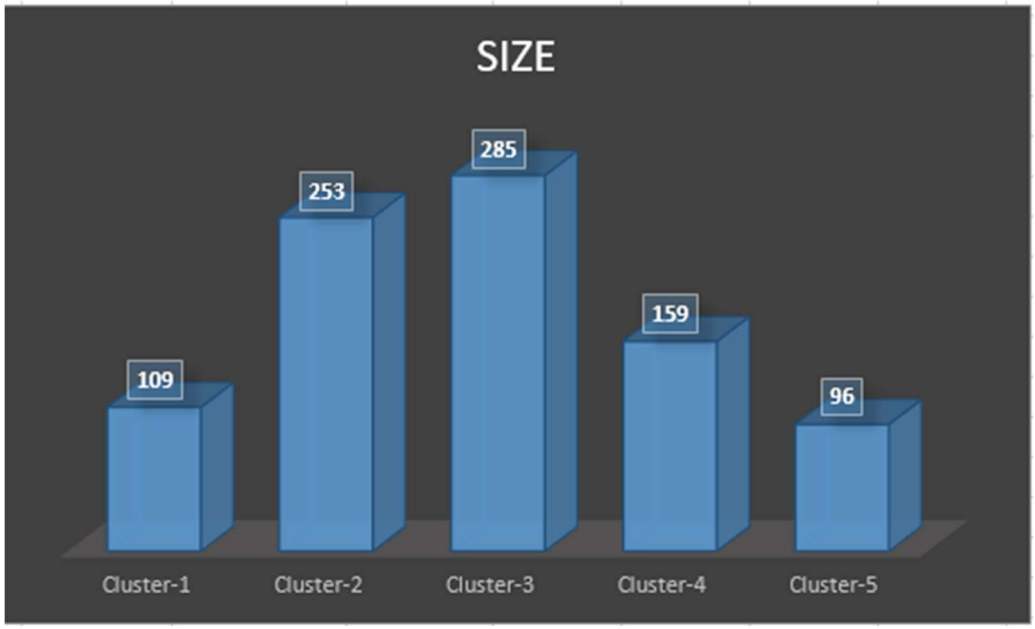
CLUSTERING:

- Clustered Customers data after eliminating outliers.
- Considered different sizes of clusters like 5,7,10,15.
- Plot the best KMeans graph to find out the most suitable number of clusters.

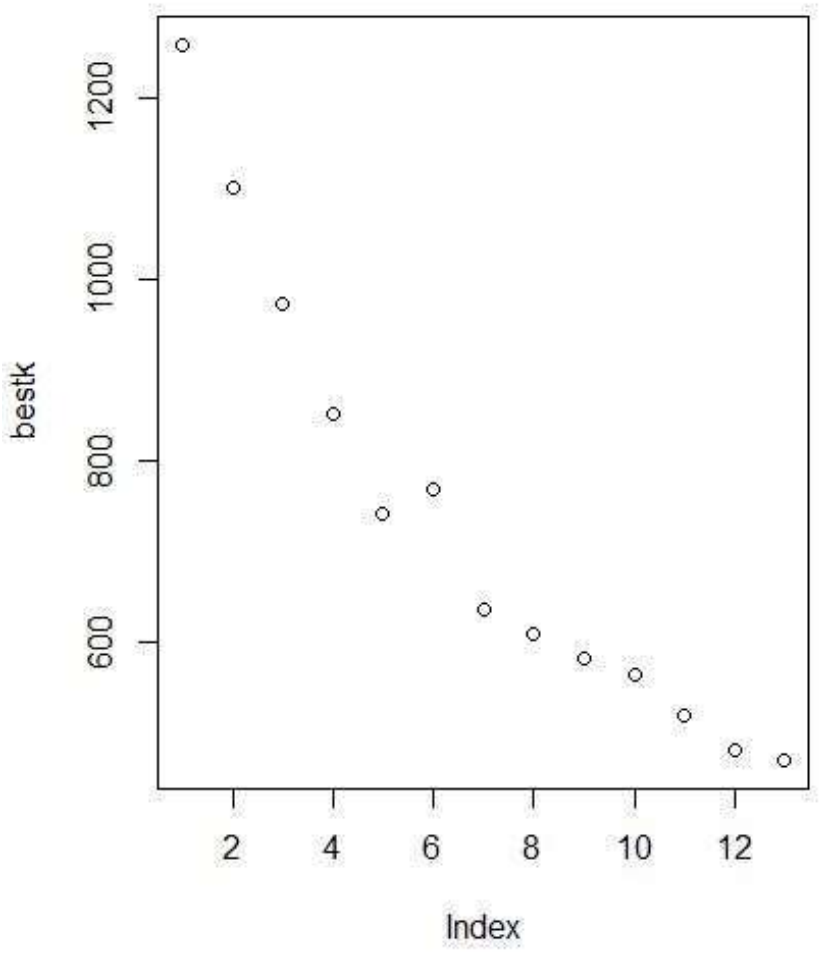
Summary of Cluster for k=5

	HighSpent	LowSpent	HighVisits	LowVisits	HighItems	LowItems
Center-1	57.43934	-57.43934	8.266055	-1.926606	103.44037	-43.72477
Center-2	60.32072	-60.32072	4.6917	-2.541502	49.71146	-14.93281
Center-3	561.51923	-159.12839	6.091228	-2.449123	65.67719	-23.87719
Center-4	133.24029	-133.24029	5.937107	-2.515723	64.54717	-23.00629
Center-5	275.52654	-275.52654	6.666667	-2.53125	62.38542	-22.11458

GRAPH THAT SHOWS CLUSTERS WITH SIZES



GRAPH THAT ALLOWS US TO CHOOSE BEST K-VALUE



CLUSTER ANALYSIS AND CLUSTER PROFILING

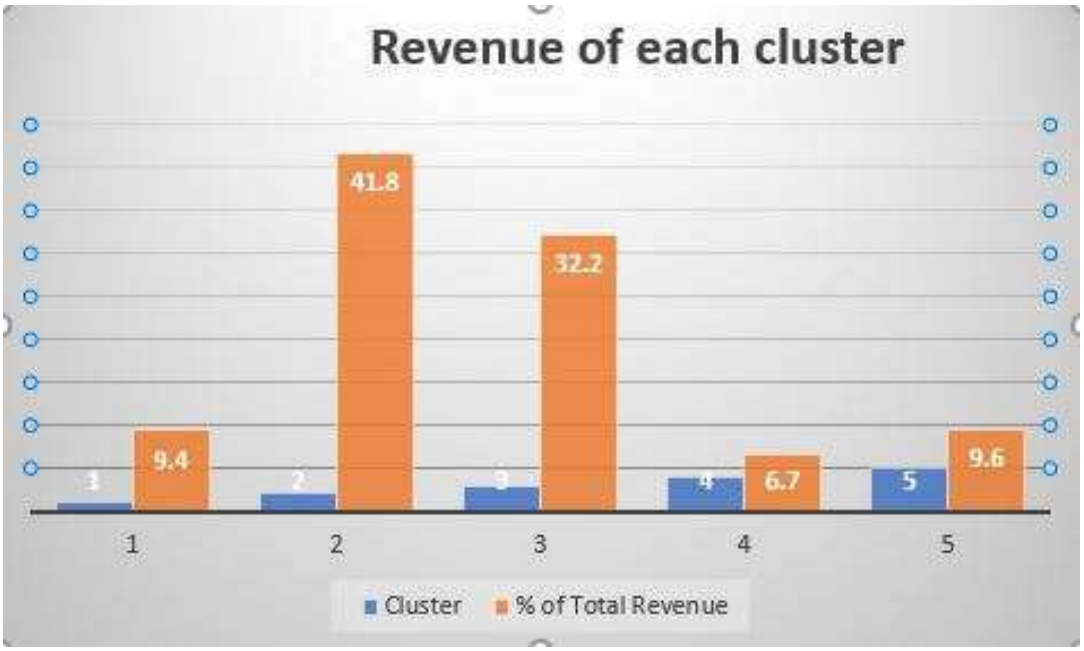
Considering Average Revenue,Visits,Items for cluster:

Cluster	Revenue	Visits	Items
1	146	9	109
2	646	9	94
3	497	4	49
4	104	3	50
5	149	9	56

Revenue, Visits and Items Analysis of Clusters:

→From above graph average number of times a customer visits a store is 9.
→Cluster 3 is best of all clusters as it has minimum products and generates customers -- belonging to this cluster generate high revenue of 32%
→Customers belonging to cluster2 generate high revenue of 32%.
→Cluster 1 has highest items and generates low revenue of just 9%
→Customers in cluster4 visit less number of times but generate a revenue of 6%
→Customers in cluster5 visit maximum number of times but generate a low revenue of 9%.This cluster is better than cluster1 as cluster5 has less items and generates same revenue as cluster1

Cluster	% of Total Revenue
1	9.4
2	41.8
3	32.2
4	6.7
5	9.6



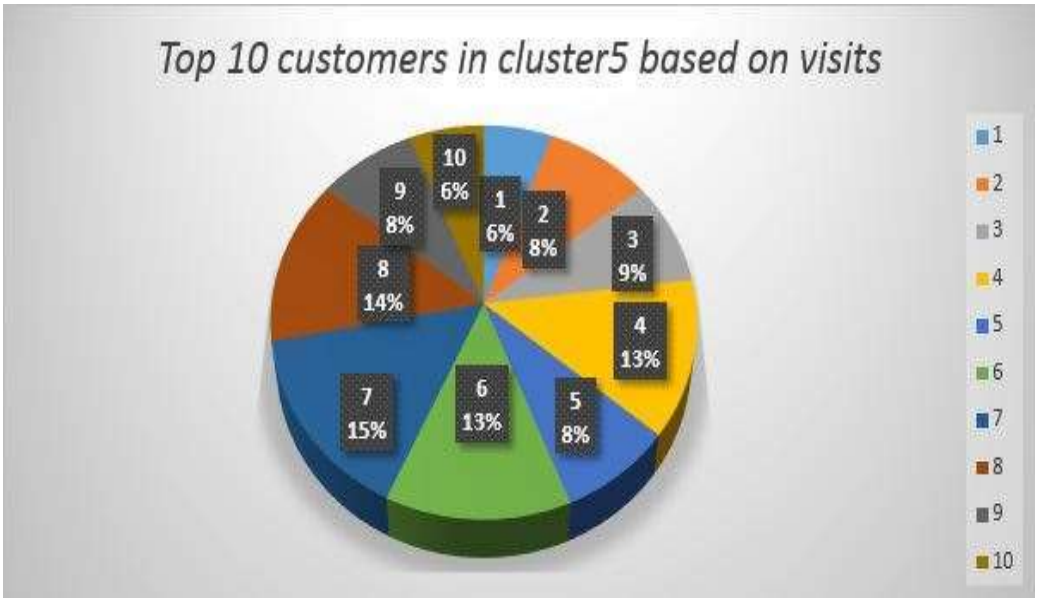
Cluster 5:

Below is the table and graph that shows top 10 customers bases on revenue

Customer_id	HighSpent
22288093	534.4694
52620857	506.2395
39349752	458.6869
40044841	458.0504
21232652	448.9807
36536337	441.9849
22190870	439.8821
30136117	417.4301
36787172	406.1552
40169314	403.1768

Below is the table and graph that shows top 10 customers bases on visits

Customer_id	HighVisits
25763791	21
39349752	17
39451170	17
58299728	17
36536337	16
58642237	16
68449510	16
62341505	16
36787172	15
28172236	15



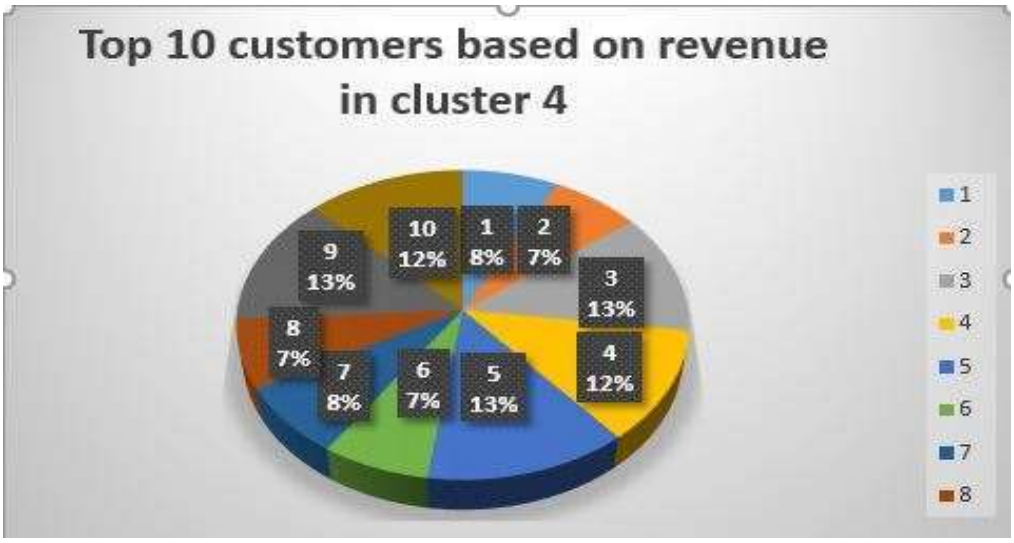
Below is the table and graph that shows top 10 customers bases on items

Customer_id	HighItems
37380276	105
23791664	94
22114139	92
53543082	87
58639848	86
22190870	84
25763791	83
59504641	83
36557632	82
39451170	80

Cluster 4:

Below is the table and graph that shows top 10 customers bases on revenue

Customer_id	HighSpent
22288093	534.4694
20811625	433.4836
37054837	427.8298
34940104	424.2132
39446850	413.5109
20842680	411.2363
22328722	408.4052
21344670	407.4151
39747934	406.6215
36783305	393.661



Below is the table and graph that shows top 10 customers bases on visits

Customer_id	HighVisits
22288093	10
51918167	8
59506824	8
59921275	7
59016293	7
58721265	7
60194995	6
61349461	6
51688908	6

62327777	6

Below is the table and graph that shows top 10 customers bases on items

Customer_id	HighItems
58892513	98
20784526	93
39937958	93
59803565	92
60415522	91
54014583	91
59564003	91
61546680	90
59788688	89
53553936	88

Below is the table and graph that shows top 10 customers bases on revenue

Customer_ID	HighSpend
22562109	976.0458
20693643	969.9541
39937223	963.4561
22286348	926.7983
36338607	923.0742
38988153	903.3299
22206723	898.9956
21352906	858.6081
39317622	857.4786
40047305	851.2233

Below is the table and graph that shows top 10 customers bases on visits

Customer_ID	HighVisits
58642266	11
21123217	10
20788623	10
22288611	10
58338429	10
61353661	10
29926497	9
54013758	9
59579364	9
51829265	8

Below is the table and graph that shows top 10 customers bases on items

Customer_ID	HighItems
39655640	90
58889706	85
59920039	83
38983543	83
59607596	82
31682231	81
58474854	81
41509991	81
58645662	80
59106914	80

Below is the table and graph that shows top 10 customers bases on revenue

Customer_ID	HighSpend
22301160	1027.703
22283251	1002.947
20943825	980.5306
22283257	939.6398
23722085	929.7554
36164600	926.3648
20739146	924.1859
39649111	921.616
21592137	851.1012
21561441	797.0326

Below is the table and graph that shows top 10 customers bases on visits

Customer_ID	HighVisits
20802419	21
40174216	21
21123450	20
22283257	19
21155118	18
21577503	17
20949611	17
22208972	15
51843234	15
51841511	15

Below is the table and graph that shows top 10 customers bases on items

Customer_ID	HighItems
22207323	143
41538512	137
22300853	135
22301160	134
40174216	128
21561441	128
28741424	126
51841511	120
22848790	118
21688304	118

Below is the table and graph that shows top 10 customers bases on visits

Customer_ID	HighVisits
60641249	22
58645609	21
58642267	21
20844587	21
66145523	20
26034751	20
58266455	19
22760741	18
58896056	18
20732494	17

Below is the table and graph that shows top 10 customers bases on items

Customer_ID	HighItems
59583005	152
36172216	152
58232932	151
61290435	149
59564012	148
58638337	147
39755604	146
21566188	146
21602722	145
59489003	144


Top 10 customers based on items in cluster1
13%
8%
12%
13%
12%
12%
8%
5%
5%
12%
1
2
3
4
5
6
7
8