

### Team member

Vanisa Achakulvisut	UIN667903568	vachak2@uic.edu
Matt Carey	UIN653922201	mcarey21@uic.edu
Abhijit Zarekar	UIN676178928	azarek3@uic.edu

### Questions

**1. What is the business goal of clustering in this case study? Describe how you will use the data provided - household demographics, purchase behavior, basis-for-purchase. Which are the variables that describe purchase behavior, and those that describe basis-for-purchase? Describe your overall approach for clustering -- you do not need to talk about different clustering methods now; write about your approach for determining number of clusters, how you will evaluate alternate clustering, etc.**

#### **What is the business goal of clustering in this case study?**

→ The business goal of clustering in this case study is to look at different groups of clustering; first we will look at **purchase behavior**, then we will look at **basis for purchase**, after that we use clusters based on both of them. If we did not break these two categories down before we compared clusters we would have a higher variance with small distance between clusters. This will help us gain information about what demographic attributes that are associated with different **purchase behaviors**(purchase process) and **degrees of brand loyalty**, so that companies can effectively deploy promotion budgets.

#### **Describe how you will use the data provided - household demographics, purchase behavior, basis-for-purchase.**

Refers to Table 1-6 below.

#### **Which are the variables that describe purchase behavior, and those that describe basis-for-purchase?**

- Purchase behavior refers to the actions taken by consumers before buying a product or service. This process may include a variety of other actions.  
We define “Purchase behavior” with these following variables; No. of Brands, Brand Runs, Total Volume, Sum of volume, No. of Trans, Number of purchase transactions, Value, Avg. Price
- Basis for purchase includes the variables which all related to purchase within promotion, price category wise purchase, selling proposition wise purchase. We define “Basis for purchase” with these following variables; 'Pur\_Vol\_No\_Promo\_\_\_\_', 'Pur\_Vol\_Promo\_6\_\_', 'Pur\_Vol\_Other\_Promo\_\_', 'Pr\_Cat\_1', 'Pr\_Cat\_2', 'Pr\_Cat\_3', 'Pr\_Cat\_4', 'PropCat\_5', 'PropCat\_6', 'PropCat\_7', 'PropCat\_8', 'PropCat\_9', 'PropCat\_10', 'PropCat\_11', 'PropCat\_12', 'PropCat\_13', 'PropCat\_14', 'PropCat\_15'

**Describe your overall approach for clustering -- you do not need to talk about different clustering methods now; write about your approach for determining number of clusters, how you will evaluate alternate clustering, etc**

The goal of clustering is to perform customer segmentation in order to get business inside and underlying useful information about customers behaviors and the potential to gain market share in future. Performing the customer segmentation helps companies develop marketing campaigns and pricing strategies to extract maximum value from both high- and low-profit customers.

→ Different promotions can be designed based on market segments to target specific segments. This would result in a more cost-effective allocation of the promotion budget to different market-segments. It would help CRISA to increase brand loyalty also.

→ Purchase Behavior:

(1) percent of purchases devoted to major brands [Ex: Is a customer only devoted to brand A?]

(2) Purchases devoted to smaller brands [Ex: To reduce complexity of analysis, a “catch-all variable for percent of purchases to the smaller brands].

(3) maximum share devoted to any one brand [Ex: Derived variable that indicates].

All variable information in this dataset are described in the following summary table:

**Table1: Demographic Data**

	Variables	Description	Reasons	Treat Null Values
D E M O G R A P H I C  D A T A	SEC	Socio economic class	Social classes are powerful indicator of purchasing behavior because different social class has different spending habits	No null values
	FEH	Food Eating Habits	Food Eating Habits indicates the purchasing power i.e. vegan has less power of purchasing than people who are not vegan	Replace NULL values with “0” Indicates “Not specified”
	MT	Native Language	Native language	Replace NULL

			most likely indicates the geographic region of the household and thus relates to the socio-economic class. It is because the purchase preferences of people change depending on the cultural geography.	values with "0" Indicates "Not specified"
	SEX	Sex of homemaker	Different genders have different purchasing habits	Replace NULL values with "0" Indicates "Not specified"
	AGE	Age of homemaker	Different ages have different purchasing habits	No null values
	EDU	Education of homemaker	Education level has different purchasing behavior	Replace NULL values with "0" Indicates Not specified
	HS	Household size	Different households have different purchasing behavior	Replace NULL values with average
	CHILD	Presence of children in household	Number of children of each household indicates different needs of consumer products	No null values
	CS	Television	Advertisement via television influences the purchasing behavior of the customers	Replace NULL values with "0" Indicates Not specified
	Affluence Index	Affluence Index	Based on the possession of certain durables we can understand the degree	

			of affluence of households and thus relates to the purchasing behavior of the households.	
--	--	--	---	--

**Table2: Purchase Summary Data**

	Attributes	Description	Reasons
Purchase Summary Data	Brands	Number of brands purchased	Number of Brands purchased gives some idea about the preferences of consumers. For example, people who buy fewer brands may indicate that they are loyal to only those brands. However, there could be a number of other factors that determine what number of brands does the customer purchase.
	Brand Runs	Number of instances of consecutive purchase of brands	Repeated purchase of same brands gives idea about the brand loyalty of customers towards particular brands.
	Total Volume	Sum of volume	Volume purchased indicates the purchase behavior of the customers. Typically affluent households buy goods in larger volumes compared to poor households.
	No. of Trans	Number of purchase transactions; Multiple brands purchased in a month are counted as separate transactions	Number of transactions also gives idea about the purchase behavior. Frequent transaction indicates the purchasing power of the customer.
	Value	Sum of value	Value of the products purchased also gives idea about the purchase behavior because different types of consumers purchase different values of goods depending on factors such as income, household strength, etc.

	Trans / Brand Runs	Avg. transactions per brand run	
	Vol/Tran	Avg. volume per transaction	
	Avg. Price	Avg. price of purchase	

**Table 3: Purchase within Promotion(sum to 100%)**

	Attributes	Description	Reasons
Purchase within Promotion	Pur Vol No Promo - %	Percent of volume purchased under no-promotion	This variable determines purchase behavior because products purchased without promotions indicate consumers' preferences.
	Pur Vol Promo 6 %	Percent of volume purchased under Promotion Code 6	This variable does most likely influence the purchase behavior of the customers. Typically customers tend to purchase products under promotion.
	Pur Vol Other Promo %	Percent of volume purchased under other promotions	This variable indicates the purchase behavior as customers tend to buy products under promotion.

**Table4: Brand wise purchase (sum to 100%)**

	Attributes	Description	Reasons
Brand wise purchase	Br. Cd. (57, 144), 55, 272, 286, 24, 481, 352, 5 and 999 (others)	Percent of volume purchased of the brand	This variable indicates the degree of brand loyalty towards the mentioned brands which again influences the purchase behavior of customers.

Following are the variables that describe basis of purchase:

**Table 5: Price category wise purchase (sum to 100%)**

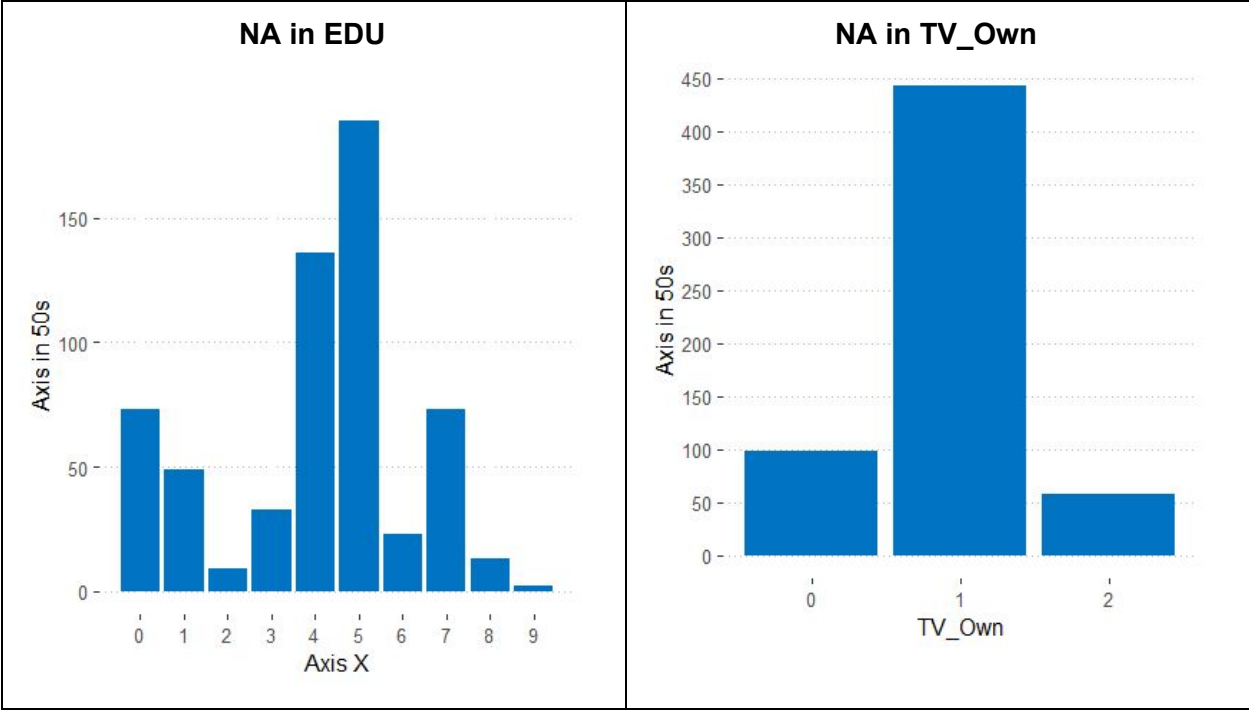
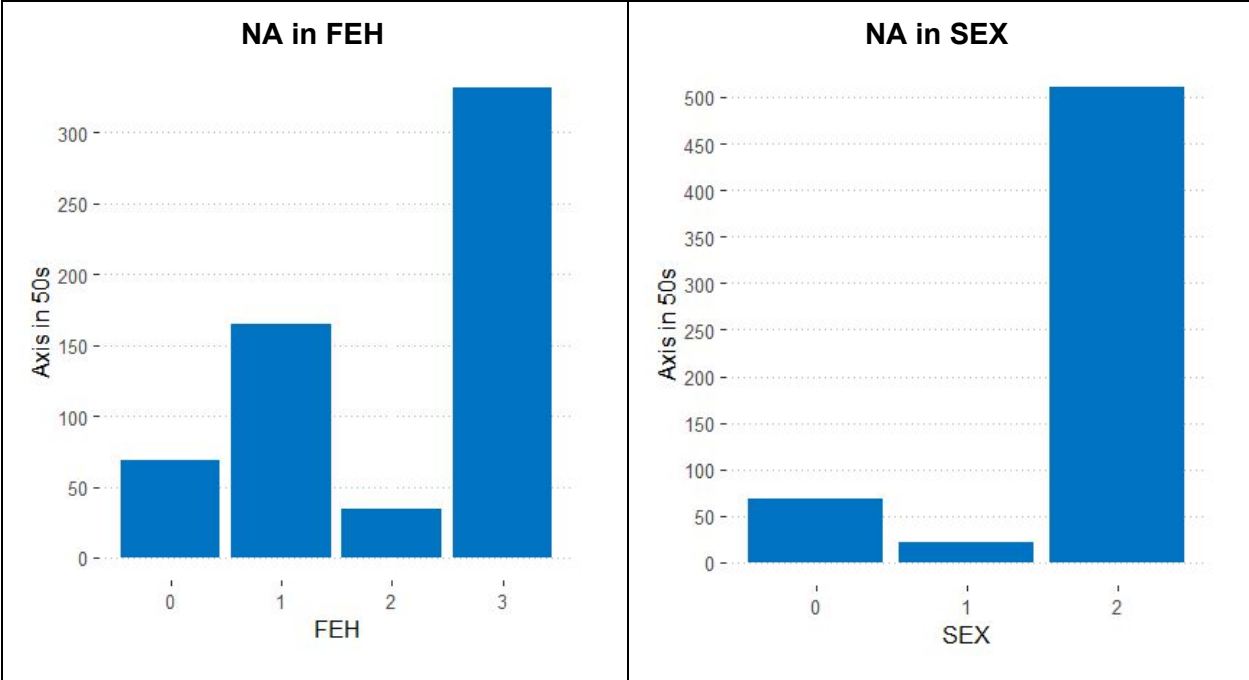
	Attributes	Description	Reasons
Price category wise purchase	Price Cat 1 to 4	Percent of volume purchased under the price category	This variable indicates the basis of purchase because price determines the decision or preferences of customers.

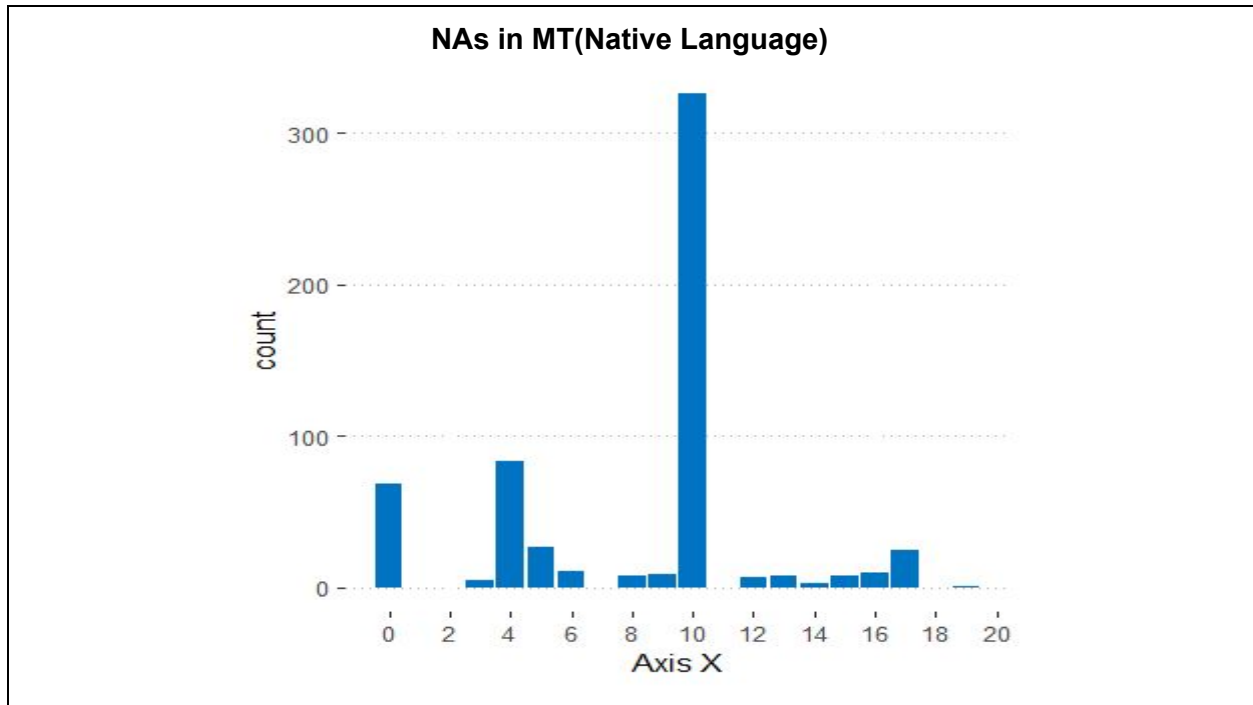
**Table 6: Selling proposition wise purchase (sum to 100%)**

	Attributes	Description	Reasons
Selling proposition wise purchase	Proposition Cat 5 to 15	Percent of volume purchased under the product proposition category	Selling proposition determines the basis of purchase because based on selling proposition customers are attracted towards purchasing a particular product. For example, a product has a selling proposition that it reduces hair fall by 90% then it is most likely to be preferred by the customers.

**2. Explore the data. Are there any missing values – how do you handle these? Summarize the households in the data based on demographic variables – use plots, tables to help your description.**

After exploring our data, we did notice that there were NA's within some of our variables. Within the 446 total NA's we saw, most of them came from demographic or household information. The information that we saw missing data from FEH, MT, SEX, EDU, HS, and CS. We are going to use this information from missing data later on in our analysis of what we are going to do for our advertising and marketing proposal. It is good for us to know information that is not given during the sampling so we can have an understanding of purchase behavior and basis of purchase before manipulating the data to replace these missing values. After transforming the data we still use the information missing as part of our analysis for brand loyalty. The NA's after being analyzed have all been replaced with a "0" value. Given the Code List breakdown of the data we know that adding a zero to any variable will not affect any other numberings, as they all begin with the lowest number of a "1". Below is a synopsis of a few of the categories that we wanted to compare to give use an idea about the number of NA values:





During our analysis we did not look up information with the household size because we are not going to be taking these numbers into account for our advertising and marketing proposal to CRISA. Below we see a breakdown of the top NA's by category:

**FEH (Food Eating Habits):** 1. Pure Vegetarian, 2. Vegetarian but eat eggs, 3. Non Vegetarian

The biggest amount comes from "3" which is not vegetarian. Could this be because the population in the region of this survey is mostly vegetarian?

**SEX (Male or Female):** 1. Male, 2. Female

The biggest amount of data missing was from Females. Could this be because females were less likely to take the survey than men?

**EDU (Education):** 1. Illiterate, 2. Literate, but no formal schooling, 3. Up to 4 years school, 4. 5-9 years of school, 5. 10-12 years of school, 6. Some college, 7. College graduate, 8. Some graduate school, 9. Graduate or professional school

Here we see the top missing information in order come from 10-12 years of school, 5-9 years of school, and college graduates (respectively). Was the survey people that were biased on having an education, or the other way around possibly?

**CS (Own a TV):** 1. Cable or broadcast TV available, 2. Unavailable

The biggest missing value is from Cable or broadcast TV available. Is there a correlation in more people just owning TV's in general?

**MT (Native Language - Mother Tongue):** 1. Assamese, 2. Bengali, 3. English, 4. Gujarati, 5. Hindi, 6. Kannada, 7. Kashmiri, 8. Konkani, 9. Malayalam, 10. Marathi, 11. Oriya, 12. Punjabi, 13. Rajasthani, 14. Sindhi, 15. Tamil, 16. Telugu, 17. Urdu, 18. Sanskrit, 19. Other

We really see a spike in both Marathi and Gurarati compared to the other native languages. This information is very beneficial for our analysis in being able to breakdown an area of where this sample survey was taken in India. Based on the knowledge of the languages



with MT and the missing values from Marathi and Gurarati. What could we take away from knowing what language was spoken?

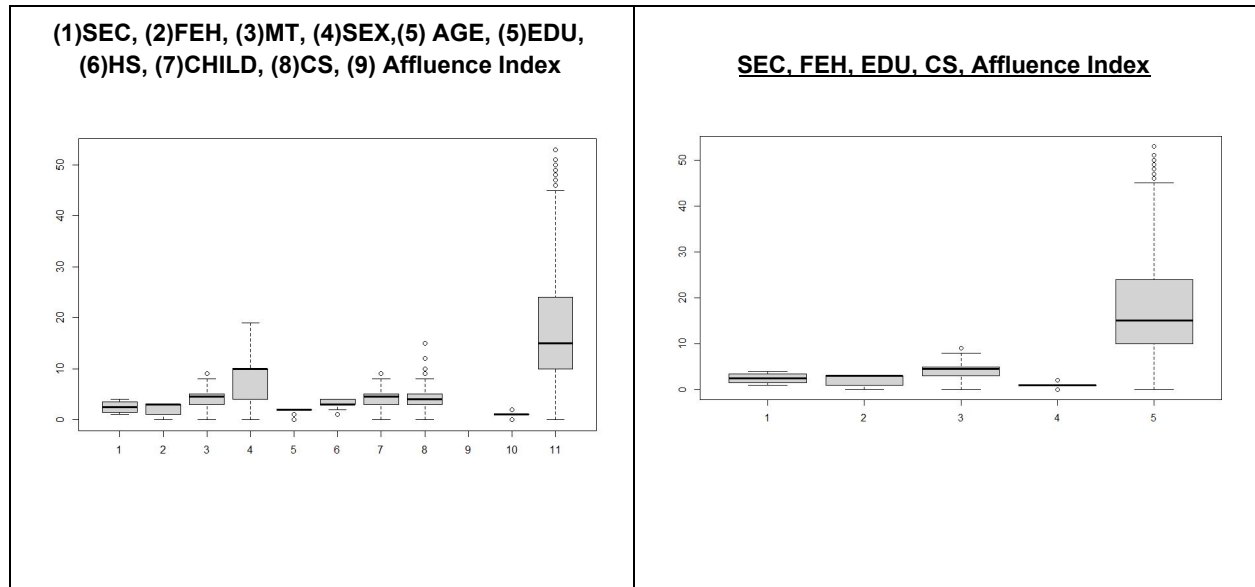
The area that we believe the survey was conducted was mainly in the state of Maharashtra, with areas from the eastern border of Gujurat, and western border of Madhya Pradesh. Knowing even the rough area of where the survey was taken helps us research more about the variables we are using and why they would contain the information they do.



**Summarize the households in the data based on demographic variables – use plots, tables to help your description.**

To describe household data we first had to figure out each variable that we have as factors. When researching the information we were given from the code list we realized that we had to look at the big picture first to see how these variables could affect our advertising and marketing proposal later on in the analysis. Depending on certain life habits or situations on the examples in our survey, these household variables could have a very significant impact on the purchase behavior and basis of purchase models we will be setting up with our clustering.

To start off our household analysis of the data we wanted to do a boxplot to see the distribution of data. After running our boxplot on the 9 different variables we see that the affluence index has a much longer range than the other variables. This is information that we need to know before doing our analysis any further so that we can understand the affluence index will have a much bigger impact against other variables that fall in a smaller statistical range of distributions.

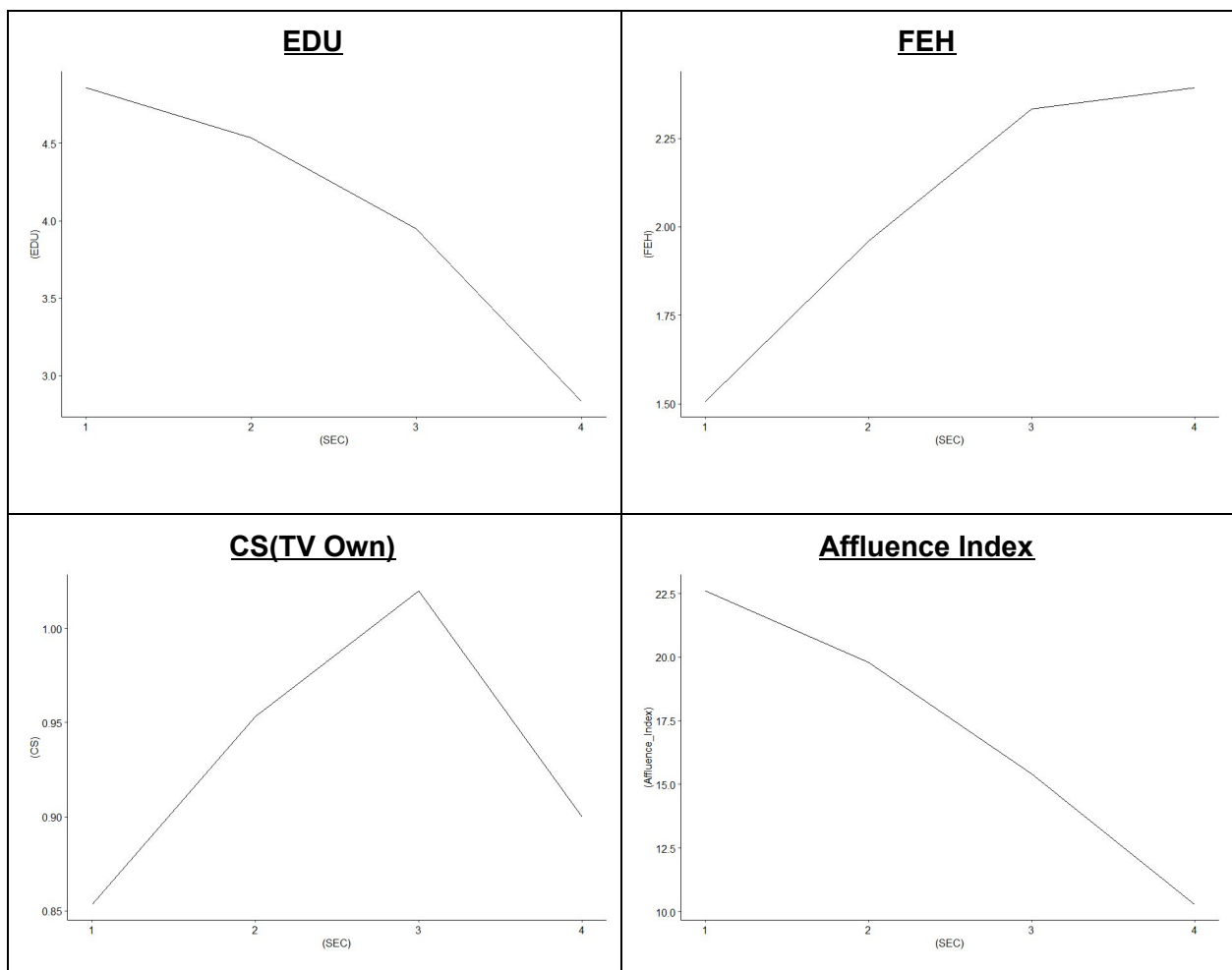


After comparing all of our different boxplots, we wanted to get more into the data and see what all the means looks like compared to specific grouped-by variables. To distinguish the difference in certain household variables we decided for our analysis to see different means grouped in comparison with social economic class, education, and affluence index. The reason we went with these 3 to draw conclusions on household goods is because of their impact on the household along with the number of criteria they had within them (SEC with 4 categories, education with 9, and affluence index calculated out to 55). The table breakdown is going to give all the means of the variables compared to the main grouped-by variables. The graphs are going to be chosen variables per section to see visually how they impact in comparison to each other. With using different amounts of categories within our 3 selections of comparisons, we are hoping to see if the amount of categories could also have an impact. .

## SEC

SEC	Member_id	FEH	MT	SEX	AGE	EDU	HS	CHILD	CS	Affluence_Index	No_of_Brands	Brand_Runs	Total_Volume	No_of_Trans	Value	Trans	Brand_Runs	Vol_Trans	Avg_Price
1.00	1138915.40	1.51	6.53	1.58	3.20	4.86	3.68	3.44	0.85	22.61	3.59	16.77	9573.21	29.99	1307.07		2.17	360.99	14.15
2.00	1122354.13	1.96	8.00	1.81	3.29	4.53	4.19	3.15	0.95	19.79	4.02	17.75	11752.70	32.09	1402.61		2.36	397.23	12.53
3.00	1098155.40	2.33	9.14	1.82	3.28	3.95	4.15	3.16	1.02	15.41	3.58	15.37	12717.80	32.09	1411.31		2.70	422.83	11.50
4.00	1057325.53	2.39	9.05	1.74	3.08	2.83	4.74	3.18	0.90	10.28	3.35	13.11	13615.37	30.45	1228.55		3.24	479.15	9.16

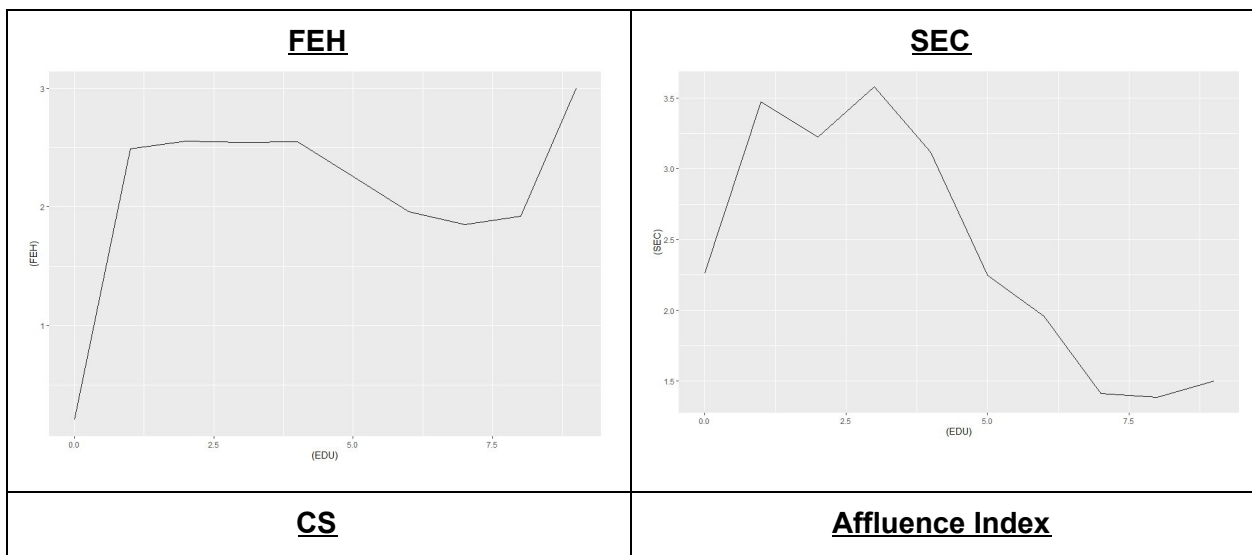
Grouping by socioeconomic class we knew was going to be an eye opener into the household data that we were going to be working with. Depending on people surveyed depending on their standings in life economically within their society has a direct impact on everything to do with their spending habits. The range of SEC is from 1-4 is the economical class level ranging from 1 = high to 4 = low. The biggest impact that we can see is on education, affluence index, and average price. As social economic class gets lower the amount of education drastically decreases from 10-12 years of school to roughly 4 years of schooling. We also see that the affluence index goes from a 22.6 to 10.3. This means that as social economic class goes down, out of the 68 items in durables, households went from owning on average 23 of those items to just around 10. The average price goes down as the economic level decreases from \$14.15 to \$9.16. Below are 4 graphs showing the social economic class (left to right, high to low) on the x-axis and education, food eating habits, TV ownership, and affluence index plotted on the y-axis.



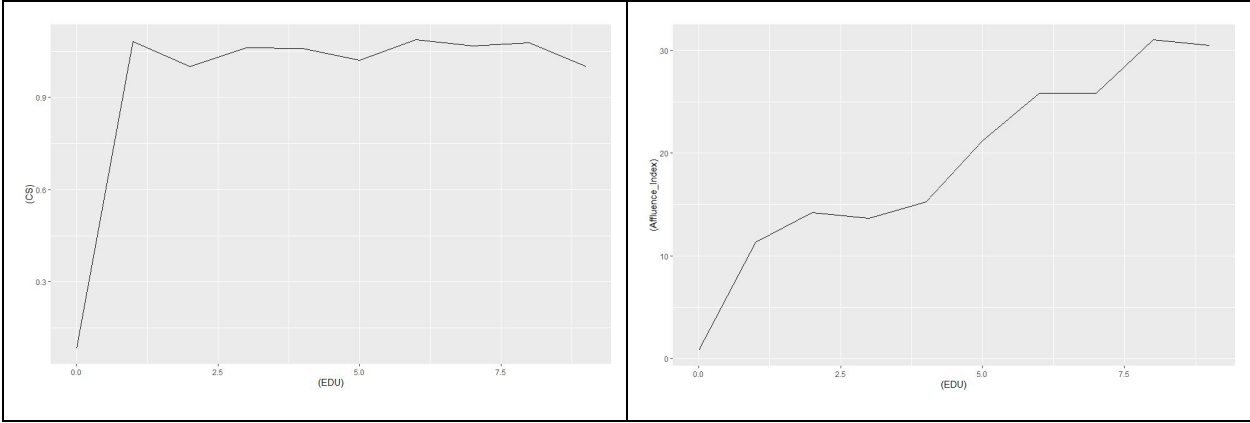
## EDU

EDU	Member_id	SEC	FEH	MT	SEX	AGE	HS	CHILD	CS	Affluence_Index	No_of_Brands	Brand_Runs	Total_Volume	No_of_Trans	Value	Trans	Brand_Runs	Vol_Tran	Avg_Price
0.00	1108126.44	2.26	0.21	0.78	0.14	2.79	0.41	4.85	0.08	0.79	2.52	7.32	5055.27	12.04	571.89		2.16	403.47	12.26
1.00	1058425.92	3.47	2.49	9.86	1.94	3.37	4.92	3.02	1.08	11.35	3.43	11.78	14236.84	29.31	1241.52		3.69	508.27	8.80
2.00	1080452.22	3.22	2.56	10.89	2.00	3.67	5.11	2.33	1.00	14.22	3.89	17.11	15325.00	35.22	1493.78		2.34	461.73	9.91
3.00	1066858.79	3.58	2.55	10.33	1.94	3.39	5.12	3.21	1.06	13.67	3.45	14.73	14613.94	32.76	1350.96		2.85	456.44	9.63
4.00	1085916.99	3.12	2.55	9.60	1.98	3.31	5.18	3.00	1.06	15.26	3.63	16.00	15584.78	34.58	1632.93		3.09	502.25	10.79
5.00	1117786.98	2.25	2.26	9.16	1.94	3.35	4.41	3.06	1.02	21.16	3.95	17.97	11531.23	34.17	1408.93		2.30	364.66	12.56
6.00	1130994.78	1.96	1.96	7.61	2.00	2.96	4.48	2.91	1.09	25.83	4.04	19.70	11896.52	37.35	1528.62		2.10	335.14	13.26
7.00	1133139.18	1.41	1.85	8.10	1.97	2.96	4.55	2.99	1.07	25.85	3.93	19.45	10509.25	34.90	1375.12		2.53	351.20	13.73
8.00	1142446.15	1.38	1.92	7.62	2.00	3.23	3.54	2.77	1.08	31.00	4.15	17.62	6735.38	28.31	1017.69		1.92	261.44	14.97
9.00	1147885.00	1.50	3.00	13.50	2.00	2.50	4.50	1.50	1.00	30.50	4.00	12.50	17375.00	21.50	2342.75		1.95	741.28	12.90

After taking a deeper look into the social economic household variables, we thought the next logical thing that would give us insight is to compare education of the homeowner. Knowing how much education someone has achieved in their life can show us relatively the type of lifestyle they are living. This is information that we can directly use to comprehend purchase habits and behaviors. The range of EDU is 1-9 ranging from being illiterate to graduate/professional school graduate (in depth descriptions in first section). We found some interesting information from looking at the averages compared to education. The biggest impact we can see are from affluence index, number of brands, and sex. Affluence again we see being a huge impact as education gets higher. We see that illiterate range from owning about 11 of the 68 durable goods to higher education averaging around 31. Next we see the number of brands slightly start to increase as education goes up. Even though this number is just slight, the average going up as education does shows that people with better education might have a direct impact on how many brands they would normally buy. Last we see that as education goes up we see more females on average compared to males. Below are 4 graphs showing the education level from (left to right, illiterate to higher education) on the x-axis and food eating habits, social economic class, TV ownership, and affluence index plotted on the y-axis.



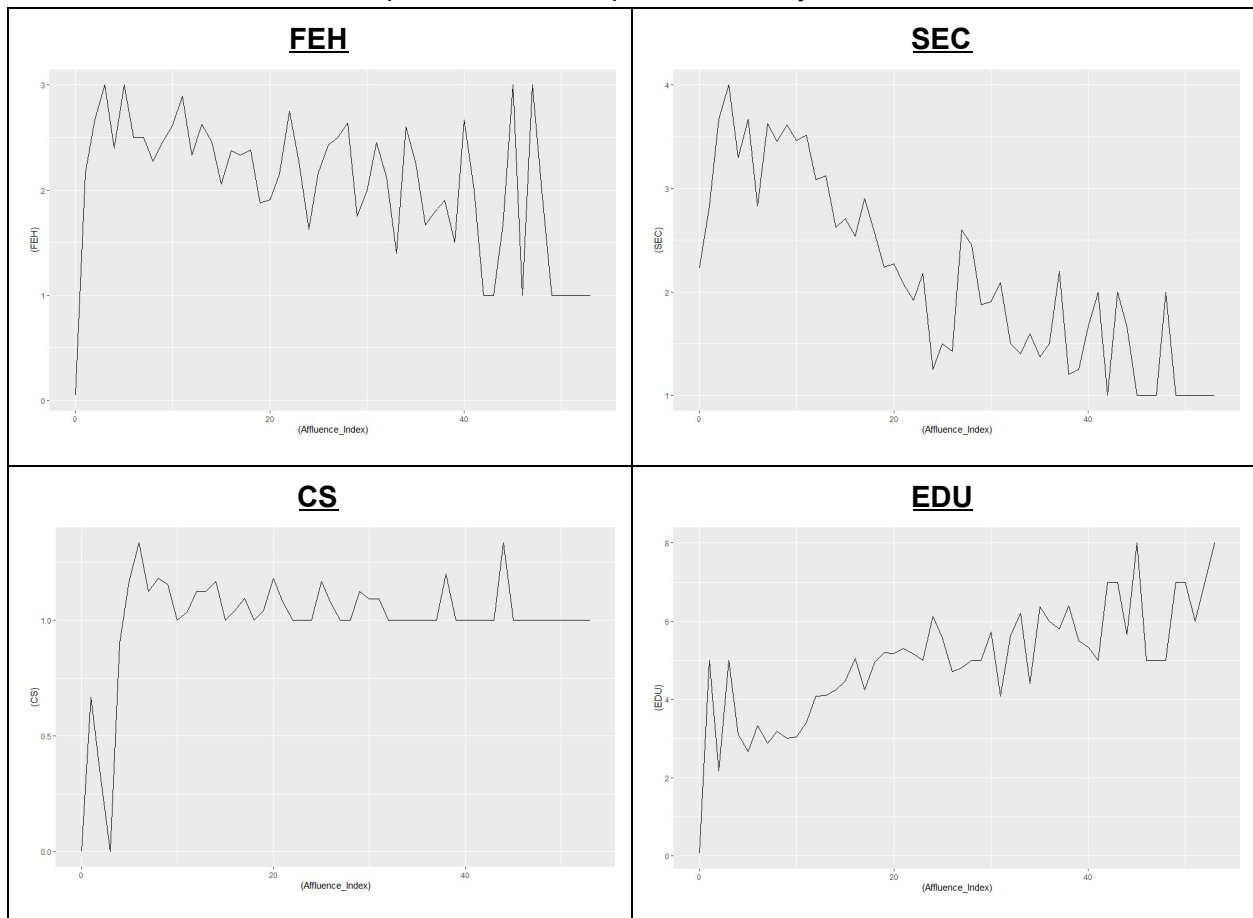




Affluence Index

Affluence_Index	Member_id	SEC	FEH	MT	SEX	AGE	EDU	HS	CHIL	CS	No_of_Brands	Brand_Runs	Total_Volume	No_of_Trans	Value	Trans	Brand_Runs	Vol_Trans	Avg	Price
0.00	1110596.23	2.23	0.04	0.14	0.03	2.70	0.06	0.07	4.97	0.00	2.45	6.67	3787.10	10.32	441.53	1.73	385.86	12.43		
1.00	1109906.67	2.83	2.17	9.17	2.00	3.50	5.00	3.17	3.17	0.67	4.00	15.17	7716.67	33.17	1143.96	3.94	247.83	14.28		
2.00	1038676.67	3.67	2.67	9.17	1.83	3.33	2.17	3.83	3.50	0.33	2.50	8.67	10383.33	25.83	770.92	4.99	398.78	7.59		
3.00	1062120.00	4.00	3.00	10.00	2.00	4.00	5.00	6.00	2.00	0.00	1.00	5.00	11900.00	25.00	1242.50	5.00	476.00	10.44		
4.00	1056539.00	3.30	2.40	10.20	2.00	3.00	3.10	4.50	2.80	0.90	3.00	9.30	13295.00	27.90	1055.75	6.13	474.68	8.03		
5.00	1053831.67	3.67	3.00	11.00	2.00	3.00	2.67	5.17	3.00	1.17	2.50	5.50	17466.67	25.33	1184.08	5.25	700.48	6.98		
6.00	1049795.00	2.83	2.50	9.17	1.83	3.17	3.33	4.17	2.67	1.33	4.17	15.33	11242.50	32.17	1112.33	2.95	369.36	10.30		
7.00	1055027.50	3.63	2.50	10.75	1.88	3.25	2.88	5.88	3.38	1.13	3.25	12.25	12965.63	31.75	1108.78	4.05	436.73	8.83		
8.00	1073468.18	3.45	2.27	10.82	1.82	2.73	3.18	4.64	2.91	1.18	3.45	13.09	14752.27	28.91	1333.27	2.83	543.10	9.52		
9.00	1070040.00	3.62	2.46	9.62	2.00	3.31	3.00	5.08	3.23	1.15	3.62	15.77	17303.46	35.00	1717.82	2.87	490.25	9.85		
10.00	1076772.31	3.46	2.62	10.04	1.96	3.31	3.04	5.58	2.62	1.00	3.58	11.65	16535.19	30.23	1530.64	3.05	574.09	9.35		
11.00	1073834.83	3.52	2.90	10.24	1.97	3.28	3.41	4.83	3.10	1.03	3.48	13.79	14920.17	31.48	1434.08	4.55	508.71	10.29		
12.00	1101912.92	3.08	2.33	10.04	2.00	3.42	4.08	4.42	3.13	1.13	3.75	16.67	11140.42	33.13	1111.59	2.22	363.96	10.27		
13.00	1081675.63	3.13	2.63	9.91	1.97	3.41	4.09	5.00	3.00	1.13	3.56	14.41	13882.50	31.22	1384.53	3.02	492.21	10.20		
14.00	1088931.67	2.63	2.46	10.50	1.92	3.50	4.25	5.50	2.71	1.17	3.54	14.58	14177.08	31.79	1509.32	3.22	469.75	11.08		
15.00	1101164.71	2.71	2.06	8.97	1.91	3.09	4.47	4.76	3.15	1.00	4.03	14.97	12533.97	32.12	1409.46	2.55	416.04	11.43		
16.00	1103726.67	2.54	2.38	8.96	1.92	3.29	5.04	4.71	2.92	1.04	3.54	15.42	13506.88	31.42	1543.96	2.14	435.04	12.81		
17.00	1093340.00	2.90	2.33	9.81	2.00	3.38	4.24	4.29	3.24	1.10	3.95	17.19	14021.90	35.33	1499.58	2.44	437.39	11.20		
18.00	1101559.05	2.57	2.38	9.14	2.00	3.10	4.95	4.00	3.48	1.00	3.81	19.52	12807.14	40.71	1518.55	2.32	354.77	12.29		
19.00	1107662.80	2.24	1.88	8.48	2.00	3.00	5.20	4.68	2.88	1.04	4.12	17.80	11328.40	34.12	1344.82	2.15	372.67	12.69		
20.00	1119761.82	2.27	1.91	7.09	1.82	3.27	5.18	4.27	2.73	1.18	4.00	20.27	11802.27	32.45	1439.55	1.87	397.57	12.48		
21.00	1132684.62	2.08	2.15	7.77	2.00	3.46	5.31	4.15	2.85	1.08	3.23	15.00	11990.38	32.62	1449.32	2.38	385.86	11.77		
22.00	1113505.00	1.92	2.75	8.42	1.92	3.33	5.17	4.75	2.83	1.00	4.58	20.25	13075.83	37.67	1830.15	1.97	382.55	15.01		
23.00	1122629.09	2.18	2.27	9.45	2.00	3.09	5.00	5.00	2.64	1.00	3.55	17.73	10554.55	29.73	1114.20	2.02	362.60	11.67		
24.00	1150626.25	1.25	1.63	7.13	2.00	3.25	6.13	4.38	3.13	1.00	4.00	16.50	8809.38	31.75	1238.47	2.59	332.49	14.24		
25.00	1130621.67	1.50	2.17	8.50	2.00	3.33	5.58	5.00	3.08	1.17	3.92	21.42	13745.00	46.50	1705.45	4.16	289.21	12.91		
26.00	1140962.86	1.43	2.43	9.43	2.00	3.43	4.71	4.57	3.00	1.07	3.93	20.07	10900.71	34.43	1484.29	2.41	314.30	14.05		
27.00	1119296.00	2.60	2.50	7.80	2.00	3.20	4.80	5.60	2.90	1.00	4.40	25.70	14728.50	41.80	1868.60	1.66	460.17	12.90		
28.00	1119766.36	2.45	2.64	8.36	2.00	3.18	5.00	4.00	3.55	1.00	3.27	14.82	14018.18	31.36	1500.30	2.29	505.72	10.86		
29.00	1132671.25	1.88	1.75	9.63	1.88	3.75	5.00	3.75	3.75	1.13	4.50	20.13	9468.75	35.38	1180.53	1.79	275.21	12.89		
30.00	1123696.36	1.91	2.00	9.36	2.00	3.18	5.73	3.73	3.09		1.09		3.82	18.09	8487.27	28.91	1203.82			
31.00	1126405.45	2.09	2.45	9.27	2.00	3.45	4.09	4.45	3.27		1.09		4.91	24.82	13950.91	40.55	1722.61			
32.00	1134885.00	1.50	2.13	7.13	1.88	3.50	5.63	4.13	2.63		1.00		3.38	14.25	8803.13	29.38	1019.31			
33.00	1142540.00	1.40	1.40	11.60	2.00	3.40	6.20	3.60	3.40		1.00		3.40	20.00	7874.00	34.40	1195.17			
34.00	1106978.00	1.60	2.60	8.80	2.00	3.60	4.40	5.00	2.80		1.00		4.60	24.20	17414.00	46.40	2425.25			
35.00	1147810.00	1.38	2.25	8.13	1.88	3.88	6.38	4.75	3.13		1.00		4.38	23.00	10168.38	36.38	1420.38			
36.00	1126226.67	1.50	1.67	6.00	2.00	3.50	6.00	5.67	2.83		1.00		4.83	24.17	15872.50	42.17	2046.13			
37.00	1090960.00	2.20	1.80	7.80	2.00	3.00	5.80	7.60	3.00		1.00		4.40	19.80	17397.00	39.80	2120.90			
38.00	1143048.00	1.20	1.90	8.90	2.00	3.10	6.40	4.60	2.40		1.20		3.60	17.40	11303.50	27.40	1459.40			
39.00	1143272.50	1.25	1.50	6.75	2.00	3.50	5.50	5.00	3.25		1.00		3.75	18.50	9761.25	35.25	1569.88			
40.00	1123143.33	1.67	2.67	10.00	2.00	3.33	5.33	3.33	4.00		1.00		3.00	25.67	12241.67	46.33	1385.92			
41.00	1159950.00	2.00	2.00	10.00	2.00	3.50	5.00	4.50	3.00		1.00		6.50	24.50	8262.50	34.00	1206.75			
42.00	1155110.00	1.00	1.00	10.00	2.00	3.00	7.00	4.00	4.00		1.00		5.00	57.00	8650.00	70.00	929.00			
43.00	1150380.00	2.00	1.00	10.00	2.00	2.00	7.00	5.00	1.00		1.00		7.00	42.00	21925.00	70.00	2715.50			
44.00	1157856.67	1.67	1.67	8.67	2.00	3.67	5.67	5.33	3.67		1.33		3.67	20.67	18758.33	39.67	3105.50			
45.00	1142276.67	1.00	3.00	12.00	2.00	3.33	8.00	6.33	2.33		1.00		3.00	9.33	19341.67	23.67	2576.33			
46.00	1152050.00	1.00	1.00	4.00	2.00	4.00	5.00	4.00	4.00		1.00		6.00	35.00	9275.00	40.00	1334.75			
47.00	1147580.00	1.00	3.00	10.00	2.00	2.00	5.00	3.00	1.00		1.00		2.00	14.00	6725.00	22.00	1191.75			
48.00	1143845.00	2.00	2.00	8.00	2.00	4.00	5.00	6.00	3.00		1.00		5.00	30.00	9292.50	41.50	1282.15			
49.00	1163830.00	1.00	1.00	4.00	2.00	2.00	7.00	4.00	2.00		1.00		4.00	11.00	3450.00	14.00	331.00			
50.00	1161130.00	1.00	1.00	10.00	2.00	2.00	7.00	5.00	1.00		1.00		5.00	74.00	11615.00	123.00	2068.75			
51.00	1121050.00	1.00	1.00	4.00	2.00	2.00	6.00	10.00	3.00		1.00		8.00	56.00	22325.00	86.00	3576.50			
53.00	1061580.00	1.00	1.00	4.00	2.00	3.00	8.00	3.00	2.00		1.00		2.00	14.00	6100.00	32.00	1887.00			

Since we started our analysis we knew that the affluence index was something that was going to be interesting for us to see throughout the analysis. Affluence index is a breakdown of the amount of durable goods each person surveyed owns. The list of durable goods is 68 items that range from anything from a radio, bicycle, tractor, pressure cooker, refrigerator, and telephone (just so you can get an idea of a list of durable goods). Since this list is a much larger range of numbers ranging from the amount of personal goods someone owns in their home, we decided to breakdown the averages compared just to get insight on what we should further see from the rest of our analysis. Instead of looking for comparisons as a whole, we viewed what type of possible comparisons we can get from the numbers in each tuple. Below are graphs breaking down what the affluence index looks like on the x-axis compared to food eating habits, social economic class, TV ownership, and education plotted on the y-axis.

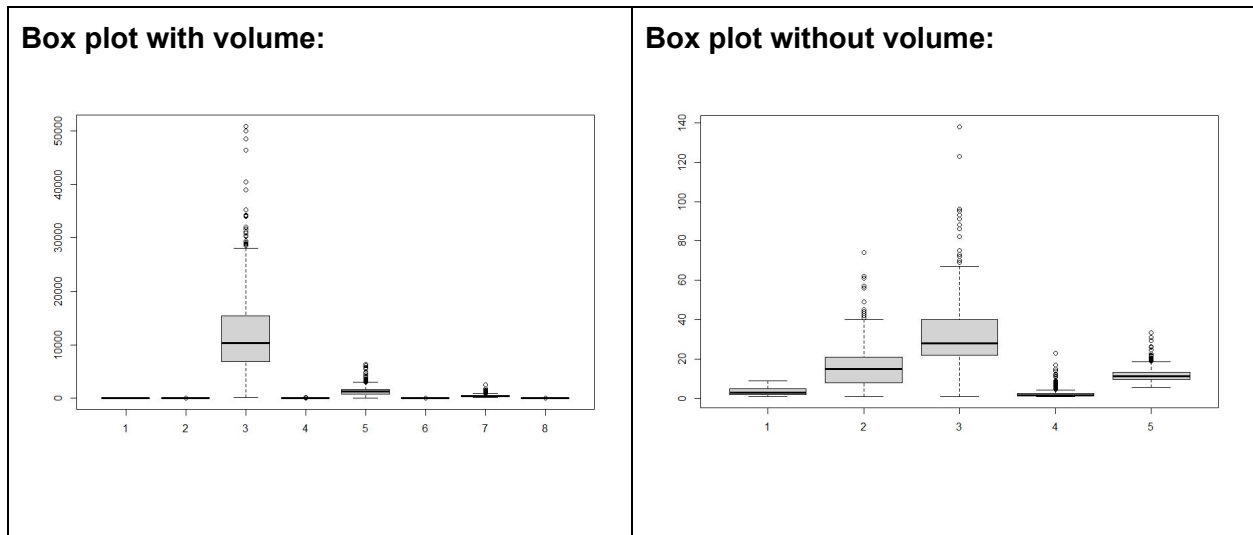


**Will you use all these variables directly, or a subset of these, and/or use any data transformations?**

Variables that could be used directly are SEC, FEH, MT, SEX, AGE, EDU, HS, CHILD, Affluence Index, No. of Brands, Brand Runs, Total Volume, No. of Trans, Value, Trans/Brand Runs, Vol/Tran, Avg. Price. These can be used directly in the evaluation parts because they are categorical variables and the values of variables are not related to other variables.

But some variables such as %Purchasing Volume on the promotion (Pur Vol No Promo - % / Pur Vol Promo 6 % / Pur Vol Other Promo %) should be evaluated as a group because all of these are proportioned and sum to 100%.

The box plot below shows the distribution of data. Labels denotes according to the (1) No. of Brands, (2) Brand Runs, (3) Total Volume, (4) No. of Trans, (5) Value, (6) Trans/Brand Runs, (7) Vol/Tran, (8) Avg. Price



As it can be seen from the above information, “total volume” has a larger range of value compared to other variables. Plus, “total volume”, “Value” and “Vol/Tran” have many outliers. Therefore, normalizing the data using scale function is required for further calculation.

**How will you evaluate brand loyalty? Describe the variables you create and use to capture different aspects of brand loyalty.**

- We measured Brand Loyalty using the following variables:
  - "Br\_\_Cd\_\_57\_\_144", "Br\_\_Cd\_\_55", "Br\_\_Cd\_\_272"
  - "Br\_\_Cd\_\_286" "Br\_\_Cd\_\_24", "Br\_\_Cd\_\_481", "Br\_\_Cd\_\_352"
  - "Br\_\_Cd\_\_5", "Others\_999"
- These variables are the Brand codes of select popular soaps. These variables show the preference of the brands by each customer measured in percentage. For example, Br\_\_Cd\_\_272 shows a value of 20% for customers with Member\_ID 1165010, that means this customer purchases Br\_\_Cd\_\_272 20% of the time.
- Therefore, Brand code with maximum percentage indicates the high preference by the customer for that particular brand.

- Hence, based on this purchase/preference measured in percent we can measure the Brand Loyalty.
- We created a new variable for Brand Loyalty called maxBr.

### 3. Use k-means clustering to identify clusters of households based on Variable index

No.	Variable name	No.	Variable name	No.	Variable name
1	Member id	21	Pur Vol Promo 6 %	41	PropCat 10
2	SEC	22	Pur Vol Other Promo %	42	PropCat 11
3	FEH	23	Br. Cd. 57, 144	43	PropCat 12
4	MT	24	Br. Cd. 55	44	PropCat 13
5	SEX	25	Br. Cd. 272	45	PropCat 14
6	AGE	26	Br. Cd. 286	46	PropCat 15
7	EDU	27	Br. Cd. 24		
8	HS	28	Br. Cd. 481		
9	CHILD	29	Br. Cd. 352		
10	CS	30	Br. Cd. 5		
11	Affluence Index	31	Others 999		
12	No. of Brands	32	Pr Cat 1		
13	Brand Runs	33	Pr Cat 2		
14	Total Volume	34	Pr Cat 3		
15	No. of Trans	35	Pr Cat 4		
16	Value	36	PropCat 5		
17	Trans / Brand Runs	37	PropCat 6		
18	Vol/Tran	38	PropCat 7		
19	Avg. Price	39	PropCat 8		
20	Pur Vol No Promo - %	40	PropCat 9		

- Describe your rationale for experimenting with different values of k.
- Evaluate the clusters – based on generic performance measures for clustering.
- Evaluate the clusters – based on the business problem and interpretation of clusters. Comment on the characteristics (demographic, brand loyalty and/or basis-for-purchase) of these clusters. This information will be used to guide the development of advertising and promotional campaigns.

a. The variables that describe purchase behavior (including brand loyalty).

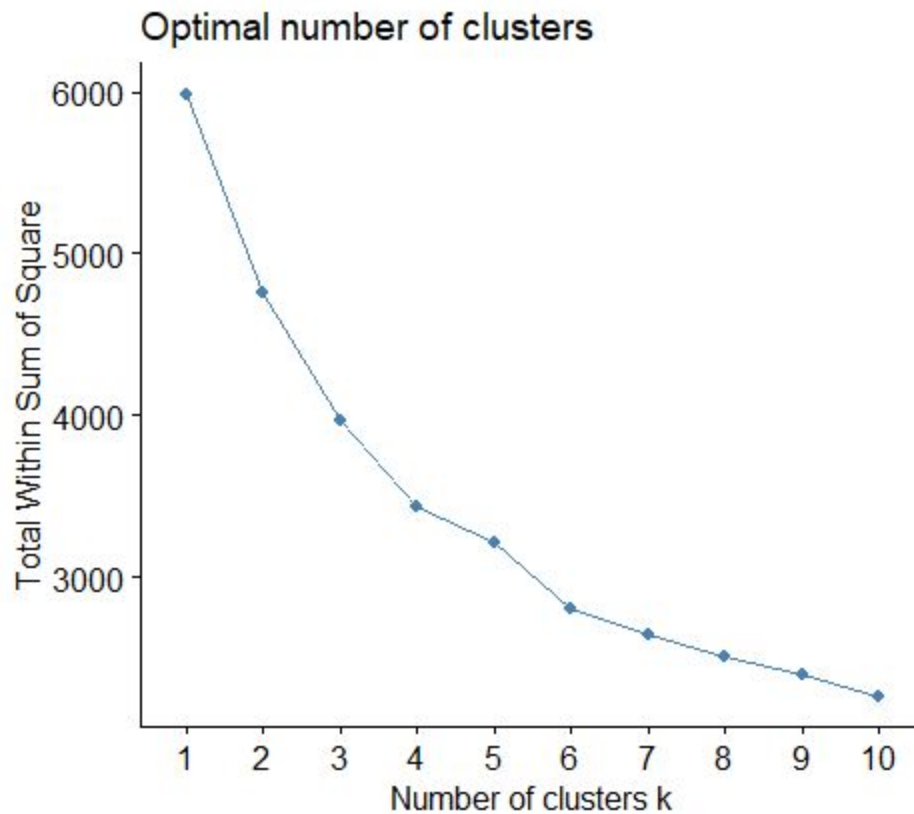
[Variables: #brands, brand runs, total volume, #transactions, value, avg. price, share to other brands, (brand loyalty)]. [Q – how do you measure brand loyalty?]

- Variables contained in PURCHASE\_BEHAVIOR are:
  - 'No\_\_of\_Brands', 'Brand\_Runs', 'Total\_Volume', 'No\_\_of\_\_Trans', 'Value', 'Trans\_\_Brand\_Runs', 'Vol\_Tran', 'Avg\_\_Price', 'maxBr', 'Others\_999'
- We measured Brand Loyalty using percentage of Brand Codes as explained [above](#).
- The new Variable measuring the Brand Loyalty is maxBr



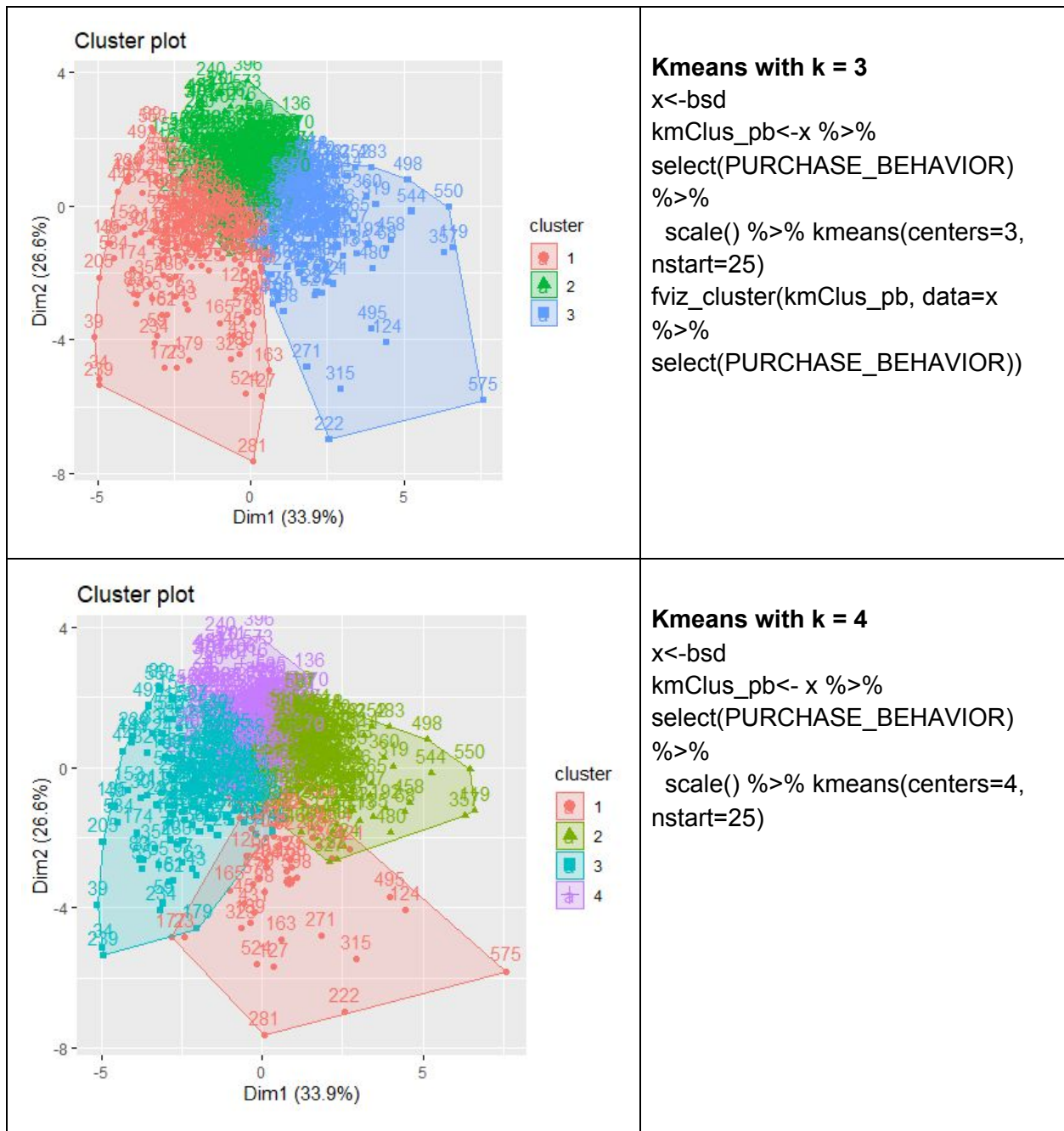
### Determining number of clusters ie k for “Purchase Behavior”

```
fviz_nbclust(xpb, kmeans, method = "wss")
```



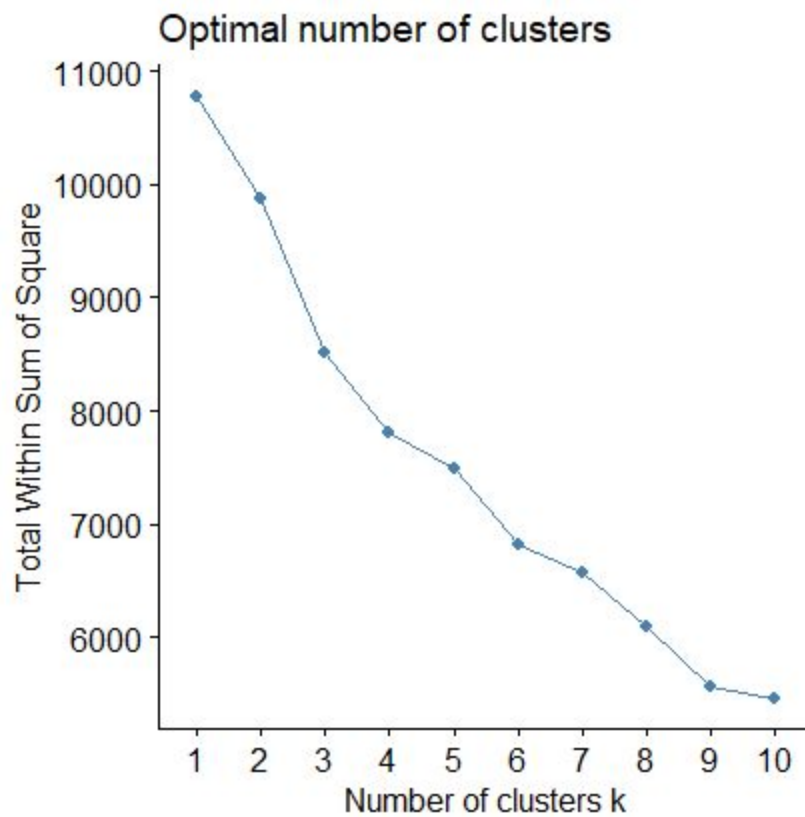
Optimal number of clusters helps to identify the suitable number of clusters by using the elbow method, the best no. of clustering (k) is on elbow sharp angle. According to this graph, we picked a reasonable k value of 3,4,6 to make a clustering model by k-mean method.

Trying K Means with different values of k = 3,4,6 ,respectively.



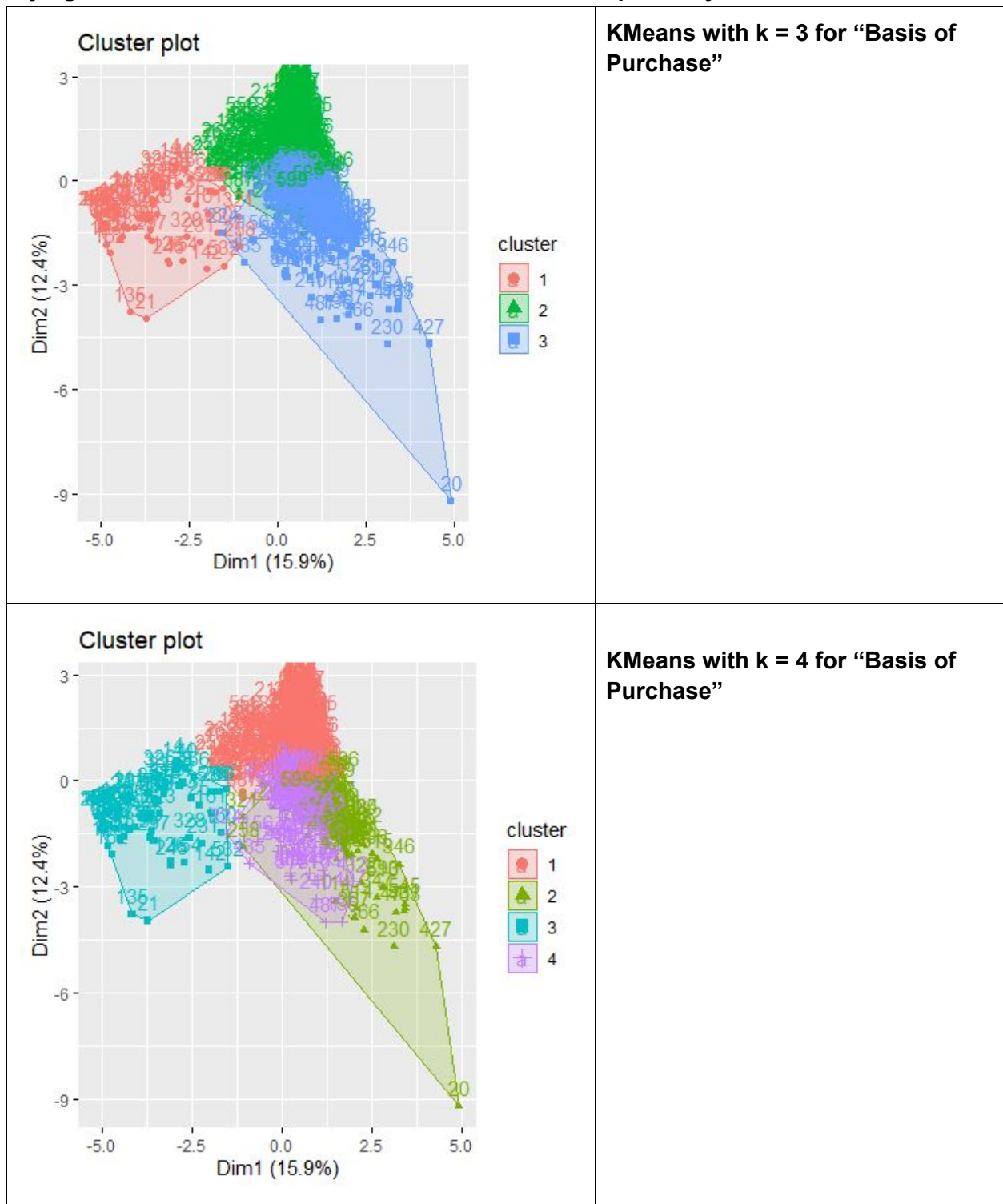


### Determining number of clusters for “Basis of Purchase”



Similarly to Basis of Purchase, we can identify the optimal number of clusters by using the elbow method. According to this graph, we picked a reasonable k value of 3,4 to make a clustering model by k-mean method.

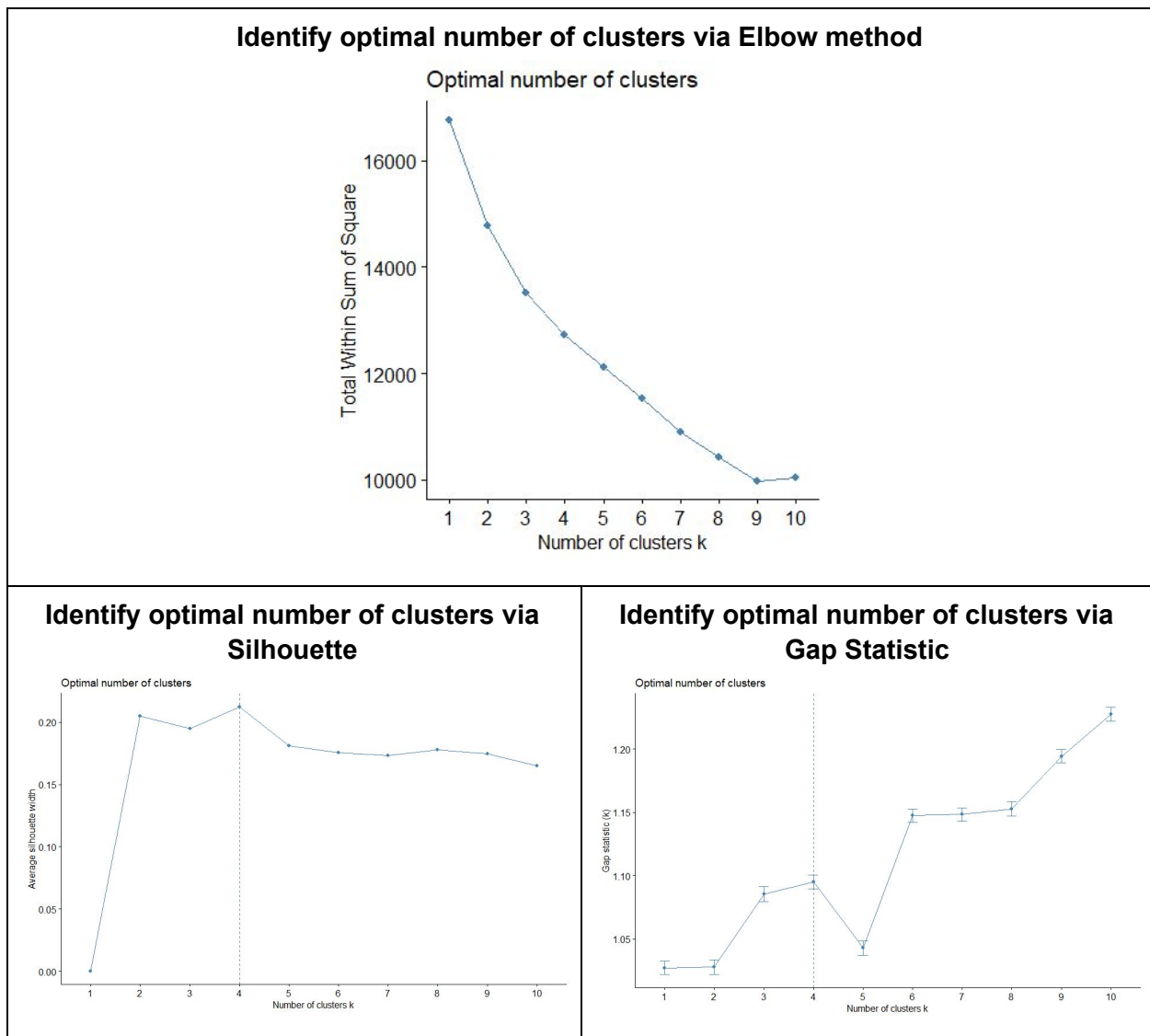
Trying K Means with different values of  $k = 3$  and  $4$ , respectively.



**c. The variables that describe both purchase behavior and basis of purchase.**

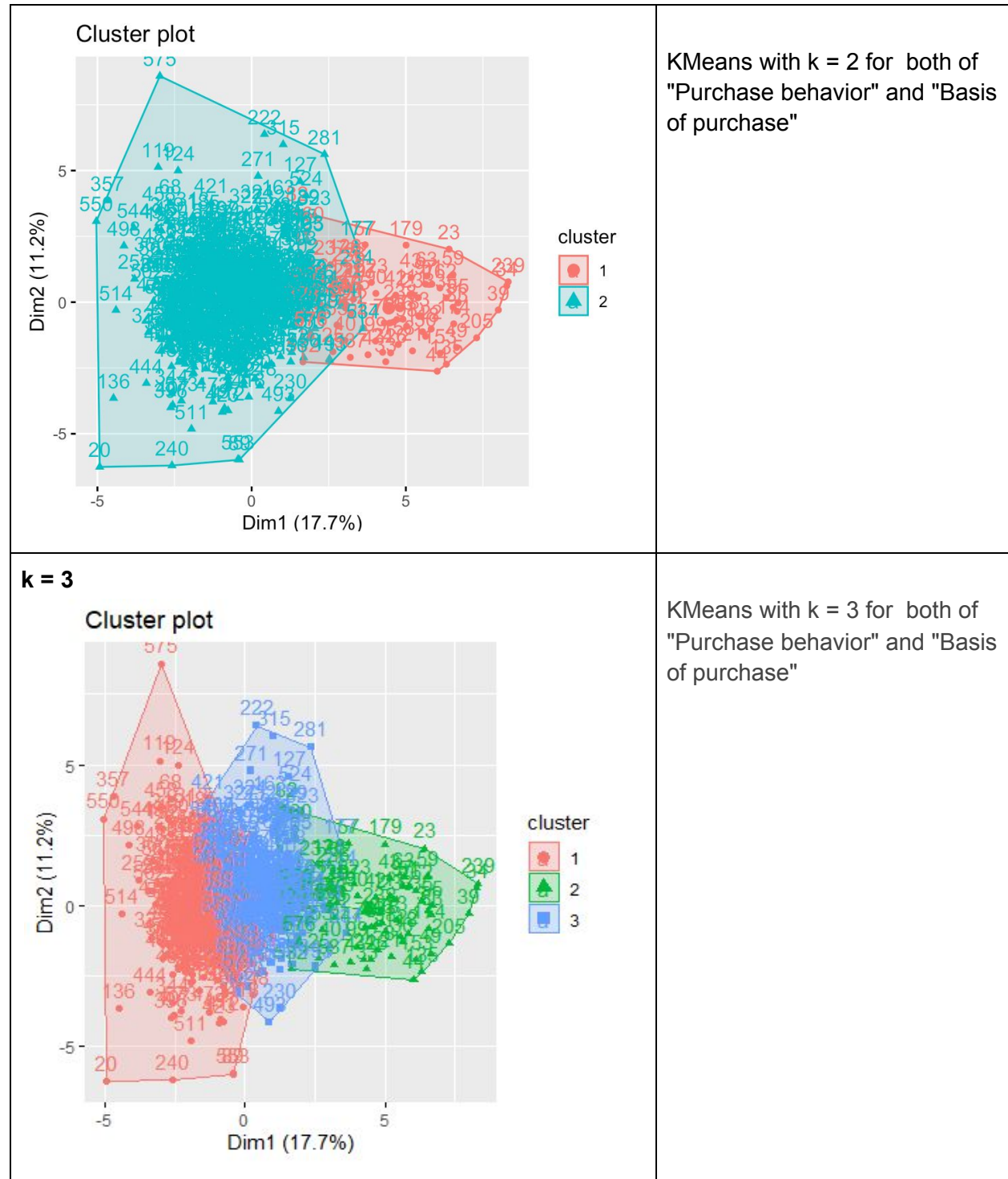
- **Variables contained in PURCHASE\_BEHAVIOR are:**
  - 'No\_\_of\_Brands', 'Brand\_Runs', 'Total\_Volume', 'No\_\_of\_\_Trans', 'Value', 'Trans\_\_Brand\_Runs', 'Vol\_Tran', 'Avg\_\_Price', 'maxBr', 'Others\_999'
- **Variables for Basis-for-purchase:**
  - **Purchase by promotions variables:** 'Pur\_Vol\_No\_Promo\_\_\_\_', 'Pur\_Vol\_Promo\_6\_\_', 'Pur\_Vol\_Other\_Promo\_\_',
  - **Price categories:** 'Pr\_Cat\_1', 'Pr\_Cat\_2', 'Pr\_Cat\_3', 'Pr\_Cat\_4',
  - **Selling propositions:** 'PropCat\_5', 'PropCat\_6', 'PropCat\_7', 'PropCat\_8', 'PropCat\_9', 'PropCat\_10', 'PropCat\_11', 'PropCat\_12', 'PropCat\_13', 'PropCat\_14', 'PropCat\_15')

**Determining number of clusters i.e. k**



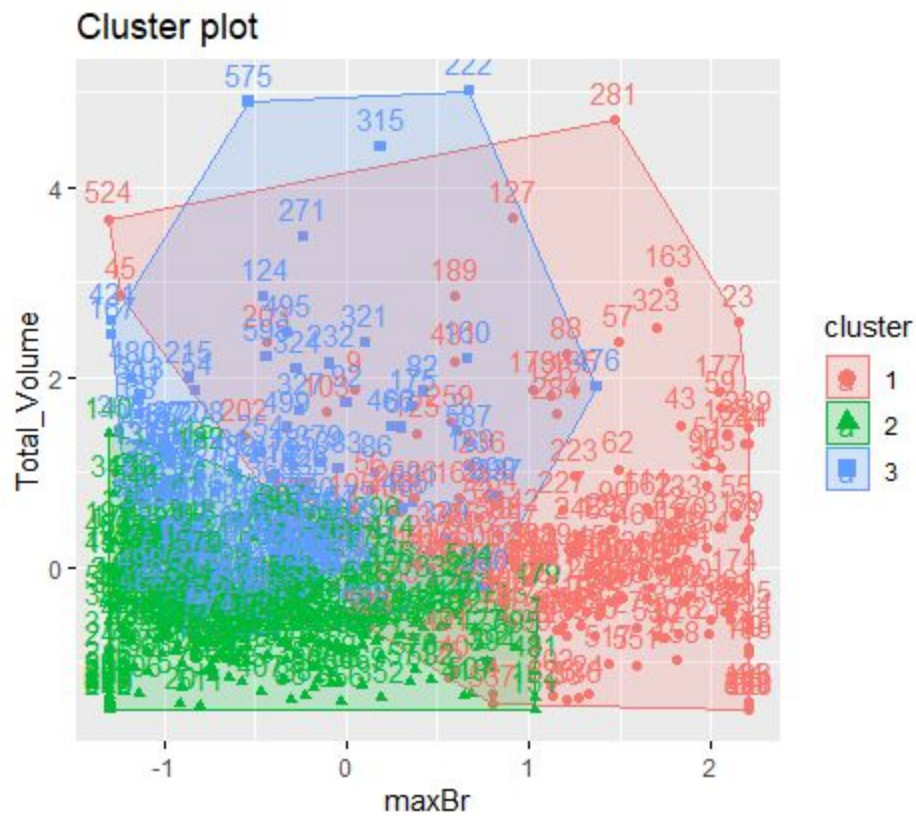
We tried different ways to find the optimal number of clusters i.e. Elbow method, Silhouette, Gap Statistic. All of these methods give a basic idea of choosing optimal numbers for clustering and the results are that the best method for finding the optimal number of clusters is different based on multiple factors including methods used and parameters for your partitioning. The elbow method looks at the total within cluster sum of squares (WSS) and minimizes it, this is the total intra-cluster variation. The gap statistic also compares the total intra cluster with a main goal of yielding the largest gap statistic. Whereas silhouette method measures the quality of the cluster by observing different k values. For our results, we saw on multiple results that 4 was the best reading in this iteration.

**K means clustering using variables that describe both "Purchase behavior" and "Basis of purchase" with k equals to 3 and 4, respectively:**





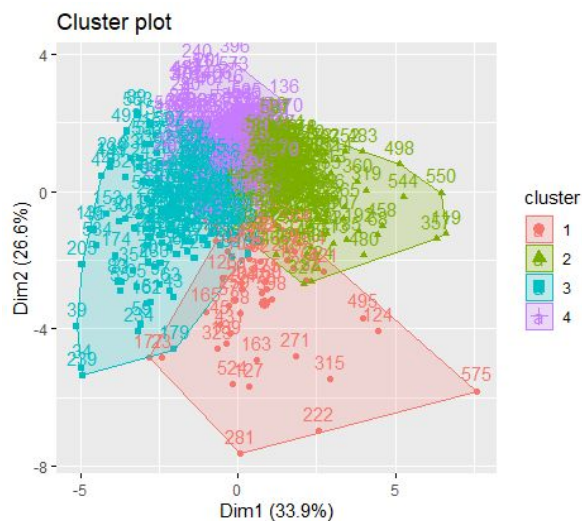
Clustering with just 2 variables:  
fviz\_cluster(kmClus\_pb, data=x %>% select(maxBr, Total\_Volume))  
# k =3



## Clustering based on Purchase Behaviour Comparison:

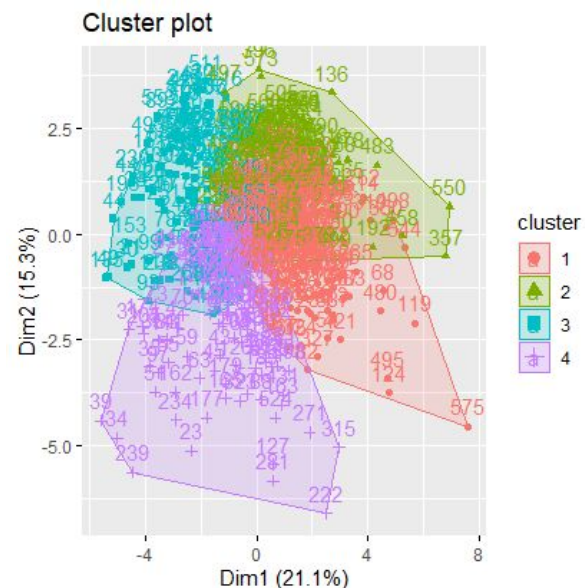
### K means clustering with k = 4 and following variables:

'No\_\_of\_Brands', 'Brand\_Runs',  
'Total\_Volume', 'No\_\_of\_\_Trans', 'Value',  
'Trans\_\_Brand\_Runs', 'Vol\_Tran',  
'Avg\_\_Price', 'maxBr', 'Others\_999'



### K means clustering with k = 4 and following variables:

'No\_\_of\_Brands', 'Brand\_Runs', 'Total\_Volume',  
'No\_\_of\_\_Trans', 'Value',  
'Trans\_\_Brand\_Runs', 'Vol\_Tran',  
'Avg\_\_Price', 'maxBr',  
'Others\_999', 'Affluence\_Index', 'FEH\_3',  
'FEH\_2', 'FEH\_1', 'SEC\_4', 'SEC\_3', 'SEC\_2',  
'TV\_1', 'TV\_2'



## DESCRIBING THE CLUSTERS:

### Description of Clusters based on PURCHASE\_BEHAVIOR and k=3:

	clusKM	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No_of_Brands	No_of_Trans	Brand_Runs	Total_Volume	Value	Trans__Brand_Runs
1	1	2.82	5.04	1.70	3.52	13.86	3.21	0.73	2.86	23.98	8.23	13349.43	1289.29	4.20
2	2	2.34	4.27	1.64	4.04	16.61	3.13	0.22	3.20	23.91	13.51	7778.10	935.56	1.97
3	3	2.41	5.23	1.92	4.60	20.99	3.35	0.24	5.14	50.01	27.18	16856.54	2015.04	1.95

- Cluster 1 shows highest Brand Loyalty at 72%.
- Clusters 2 and 3 show low Brand Loyalty
- The Volume purchased in Cluster 1 is also showing a decent figure compared to other volumes
- In case of cluster 3, the Affluence Index is higher, but the brand loyalty is similar to cluster 2
- In cluster 3, "No\_\_of\_\_Trans" are more than twice the "No\_\_of\_\_Trans" of cluster 1 and 2 each. And this is relatable to the affluence index - Affluent households usually spend frequently.

### Description of Clusters based on PURCHASE\_BEHAVIOR and k=4:

	clusKM	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No__of_Brands	No__of__Trans	Brand_Runs	Total_Volume	Value	Trans___Brand_Runs
1	1	2.70	4.71	1.65	3.52	13.89	3.11	0.72	2.96	22.57	8.19	10280.09	1003.58	3.97
2	2	2.32	4.73	1.90	4.83	21.30	3.26	0.23	5.16	46.73	26.51	13197.14	1604.91	1.85
3	3	2.83	6.70	2.00	4.20	18.80	3.48	0.40	3.54	36.87	16.02	29981.09	3186.11	2.84
4	4	2.41	4.37	1.60	3.71	15.24	3.19	0.18	2.78	22.31	12.03	7799.24	934.67	2.08

- Cluster formation with k=4 is not much different from cluster with k =3.
- Similar to k=3, Cluster 1 in here shows highest brand loyalty

### Description of Clusters based on PURCHASE\_BEHAVIOR and k=6

	clusKM	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No_of_Brands	No_of_Trans	Brand_Runs	Total_Volume	Value	Trans_Brand_Runs
1	1	2.39	4.25	1.53	3.50	14.07	3.11	0.16	2.41	19.70	10.01	6866.59	810.62	2.23
2	2	2.68	4.55	1.61	3.65	14.22	3.10	0.66	3.17	22.46	9.78	9340.65	982.73	2.53
3	3	2.89	6.66	2.00	4.18	18.43	3.45	0.36	3.34	33.89	15.23	28901.25	3018.31	2.51
4	4	3.03	5.45	1.86	3.14	11.14	3.41	0.95	2.07	24.03	2.62	16649.14	1165.36	11.05
5	5	2.24	4.49	1.85	4.81	20.40	3.23	0.22	4.66	37.47	22.02	10748.12	1361.62	1.79
6	6	2.47	5.53	1.97	4.48	23.55	3.43	0.24	5.78	64.83	34.14	19060.86	2293.06	2.09

- Cluster 4 is showing Brand Loyalty of 95% which is quite high
- Cluster 2 is showing a Brand Loyalty of 66% which is 2nd highest.
- Trans\_\_\_Brand\_Runs = No\_of\_Trans/Brand\_Runs
- We observe that Trans\_\_\_Brand\_Runs value is high for clusters which show high “brand loyalty”
- The Trans\_\_\_Brand\_Runs is high for cluster 4 compared to other Trans\_\_\_Brand\_Runs of other clusters.
- Though No\_of\_Trans are related to Affluence Index, it does not show a significant relation within Brand Loyalty.
- Overall, We can see the direct relationship between Brand\_Runs, maxBr, and Trans\_Brand\_Runs. Throughout all the different clusters from 3, 4, and 6 these had a direct reflection with brand loyalty.

### Description of Clusters based on PURCHASE\_BEHAVIOR with additional variables and k=4:

**Variables :** 'No\_\_of\_Brands', 'Brand\_Runs', 'Total\_Volume', 'No\_\_of\_\_Trans', 'Value', 'Trans\_\_\_Brand\_Runs', 'Vol\_Tran', 'Avg\_\_Price', 'maxBr', 'Others\_999', 'Affluence\_Index', 'FEH\_3', 'FEH\_2', 'FEH\_1', 'SEC\_4', 'SEC\_3', 'SEC\_2', 'TV\_1', 'TV\_2'

	clusKM	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No_of_Brands	No_of_Trans	Brand_Runs	Total_Volume	Value	Trans__Brand_Runs
1	1	2.70	4.71	0.78	1.46	4.12	2.89	0.40	2.49	14.85	7.63	5795.31	551.07	2.22
2	2	2.59	4.66	1.94	4.64	20.37	3.23	0.26	4.22	39.64	20.71	13003.93	1561.90	2.08
3	3	1.86	4.27	1.97	5.25	23.59	3.34	0.28	4.06	32.99	19.17	9791.66	1286.16	1.95
4	4	3.04	5.70	1.98	3.74	13.98	3.32	0.70	2.99	27.75	8.94	19038.84	1758.62	5.10

### BASIS\_OF\_PURCHASE

#### Description of Clusters based on BASIS\_OF\_PURCHASE and k=3:

	clusKM	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No_of_Brands	No_of_Trans	Brand_Runs	Total_Volume	Value	Trans__Brand_Runs
1	1	3.33	5.22	1.58	2.47	9.14	3.04	0.74	2.99	26.04	8.94	13538.33	947.41	5.01
2	2	2.66	4.88	1.82	4.17	16.77	3.22	0.36	3.77	31.35	15.76	13119.76	1408.31	2.37
3	3	1.89	4.37	1.67	4.47	20.64	3.27	0.24	3.67	32.89	18.49	9204.51	1374.10	2.07

#### Description of Clusters based on BASIS\_OF\_PURCHASE and k=4:

	clusKM	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No_of_Brands	No_of_Trans	Brand_Runs	Total_Volume	Value	Trans_Brand_Runs
1	1	2.59	4.81	1.81	4.18	17.15	3.23	0.37	3.82	31.76	15.92	13148.65	1427.65	2.40
2	2	3.33	5.24	1.57	2.42	9.08	3.04	0.75	2.99	25.82	8.74	13269.61	923.29	5.09
3	3	1.79	4.56	1.59	4.49	19.87	3.19	0.24	3.42	31.94	16.50	9031.75	1471.12	2.27
4	4	2.41	4.44	1.83	4.35	19.48	3.34	0.22	3.84	32.46	19.96	10268.60	1202.79	1.78

## **PURCHASE\_BEHAVIOR and BASIS\_OF\_PURCHASE**

**Description of Clusters based on both set of PURCHASE\_BEHAVIOR and BASIS\_OF\_PURCHASE with k=2:**

	clusKM	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No_of_Brands	No_of_Trans	Brand_Runs	Total_Volume	Value	Trans_Brand_Runs
1	1	3.42	5.17	1.56	2.35	8.69	3.00	0.77	2.83	24.35	7.83	12980.07	883.82	5.23
2	2	2.38	4.70	1.76	4.27	18.16	3.24	0.32	3.75	32.08	16.83	11769.50	1399.24	2.26

**Description of Clusters based on both set of PURCHASE\_BEHAVIOR and BASIS\_OF\_PURCHASE with k=3:**

	clusKM	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No_of_Brands	No_of_Trans	Brand_Runs	Total_Volume	Value	Trans_Brand_Runs
1	1	2.74	4.90	1.78	3.98	15.69	3.22	0.40	3.51	27.89	13.38	13105.01	1375.81	2.46
2	2	1.93	4.45	1.74	4.63	21.17	3.27	0.21	4.04	37.11	21.00	10074.75	1425.79	2.02
3	3	3.40	5.18	1.56	2.38	8.89	3.01	0.77	2.86	24.88	8.23	13118.70	898.42	5.19

**For each clustering in Q3 and in Q4 below:**

(i) Describe your rationale for experimenting with different values of k.  
(ii) Evaluate the clusters – based on generic performance measures for clustering.  
(iii) Evaluate the clusters – based on the business problem and interpretation of clusters.  
**Comment on the characteristics (demographic, brand loyalty and/or basis-for-purchase) of these clusters. This information will be used to guide the development of advertising and promotional campaigns.**

**(i) Describe your rationale for experimenting with different values of k in kmeans clustering.**

- We experiment with different values of k based on:
  - Variance in cluster(**WSS**)
  - Distance between the clusters
- We looked for:
  - Low Variance in clusters ie. low wss value
  - Large distance between the clusters
- We get the Within-cluster Sum of Squares (WSS) value using fviz\_nbclust() and using “**Elbow method**” we determine the optimal number of clusters.
- In Elbow method as shown in [this graph](#) we observe the sharp bend in the graph. The point with significant bend indicates the optimal number of clusters given on X-axis.
- We tried with different values of k based on the Elbow method selecting the points which show significant bend.

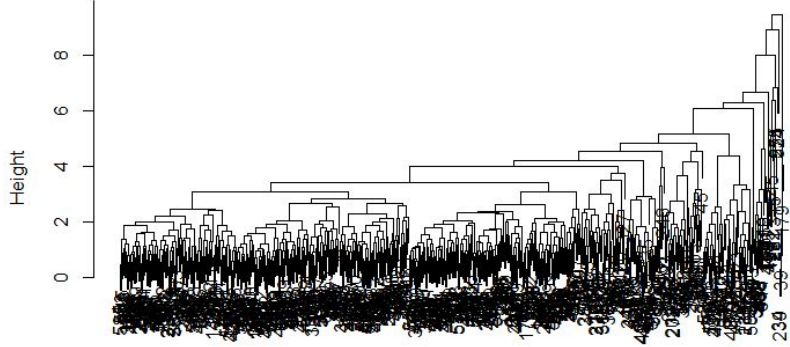
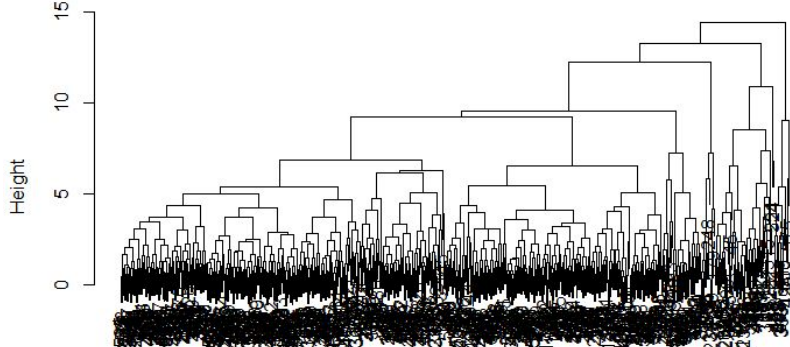
**(ii) Evaluate the clusters – based on generic performance measures for clustering.**

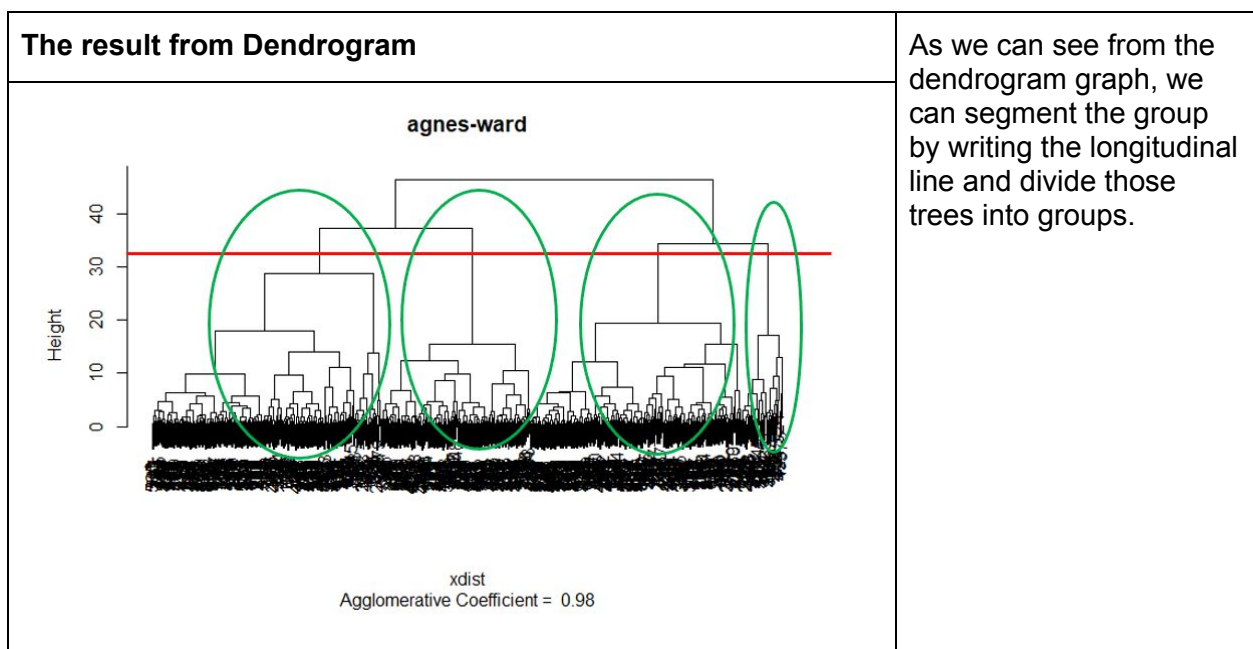
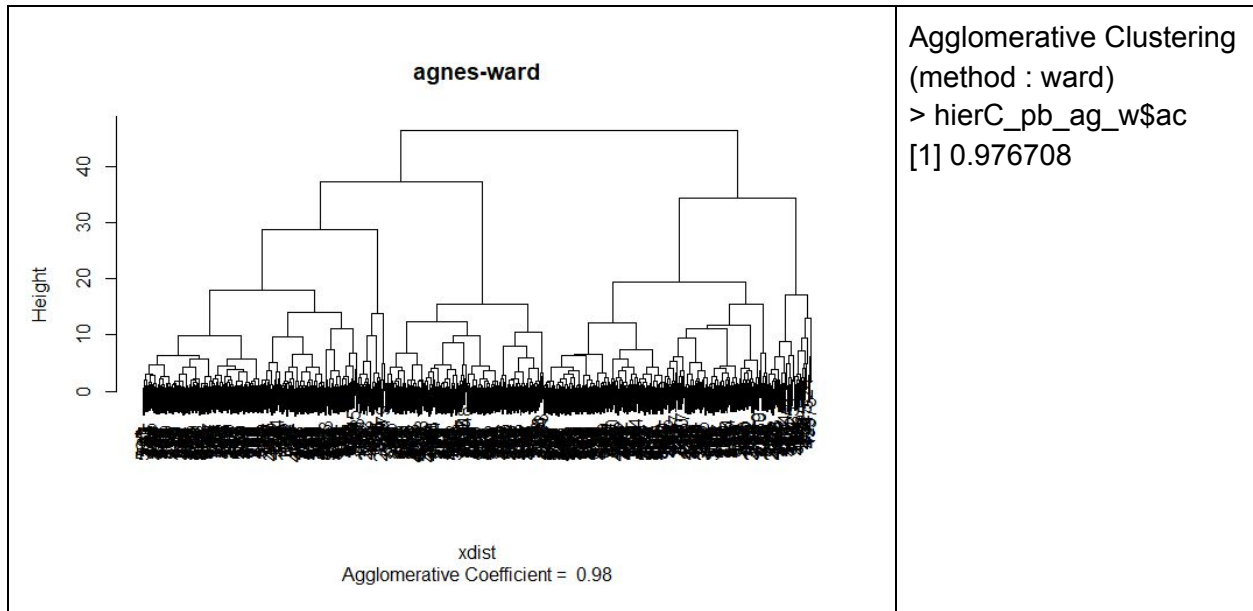
- Evaluation of the clusters is mainly done using:
  - (between\_SS / total\_SS ) value
  - We get this value after implementing kmeans algorithm.

- A higher value of  $(\text{between\_SS} / \text{total\_SS})$  indicates that there is significant separation between the clusters. The maximum value of  $(\text{between\_SS} / \text{total\_SS})$  is 1.
- This ratio gives the distance between the clusters (the larger the better)
- For clustering based on PURCHASE\_BEHAVIOR with  $k = 6$  we get
  - $(\text{between\_SS} / \text{total\_SS} = 53.2 \%)$
- However, the values are low for  $k = 3, 4$ .
- For clustering based on PURCHASE\_BEHAVIOR with additional set of variables with  $k = 4$  we get
  - $(\text{between\_SS} / \text{total\_SS} = 31.6 \%)$
  - It value does not improve even if the number of clusters are increased to 8 (44%)
- Therefore, [clustering with k =6](#) based on PURCHASE\_BEHAVIOR (original set of variables) gives better clustering.

4. Try two other clustering methods (*for a 2-person team, try one other method*) for the questions above - from agglomerative clustering, k-medoids, kernel-k-means, and DBSCAN clustering. Show how you experiment with different parameter values for the different techniques, and how these affect the clusters obtained.

### Agglomerative Clustering

<p>Dendrogram of <code>agnes(x = xdist, method = "average")</code></p>  <p>xdist Agglomerative Coefficient = 0.89</p>	<p>Agglomerative Clustering (method : average) &gt; hierC_pb_ag_a\$ac [1] 0.8871901</p>
<p><code>agnes-complete</code></p>  <p>xdist Agglomerative Coefficient = 0.93</p>	<p>Agglomerative Clustering (method : complete) &gt; hierC_pb_ag_c\$ac [1] 0.9250937</p>

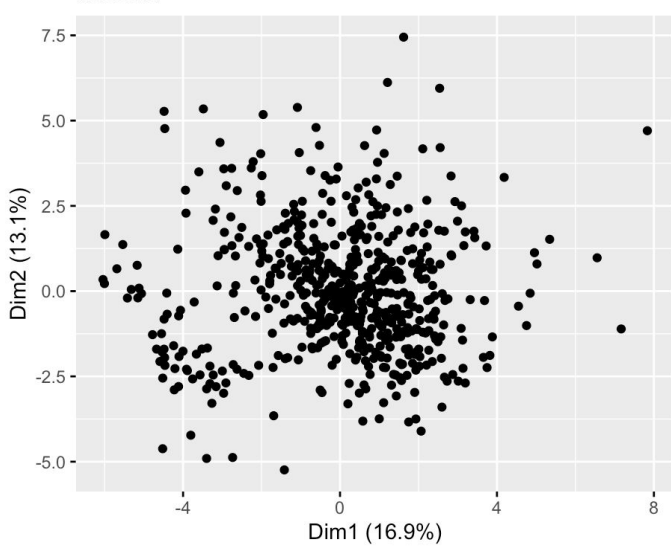
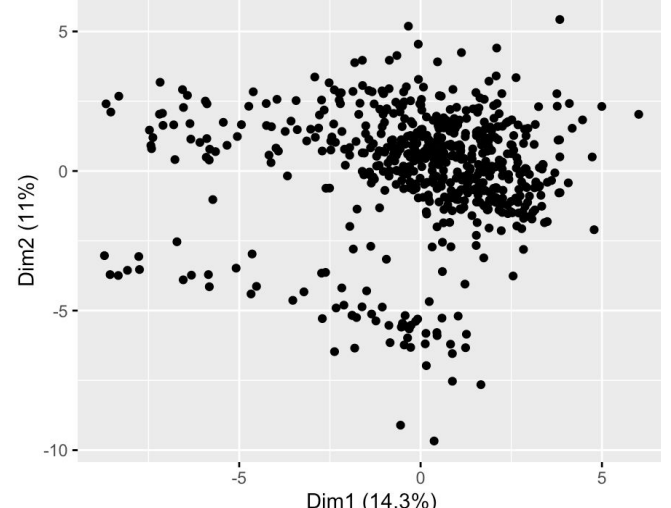


## DBSCAN

DBSCAN Parameters for clustering:

1. Epsilon (eps): It is defined as the maximum distance between two points to be considered as neighboring points (belonging to the same cluster).
2. Minimum Points (min\_samples or minPts): This defines the minimum number of neighboring points that a given point needs to be considered a core data point which includes the point itself.

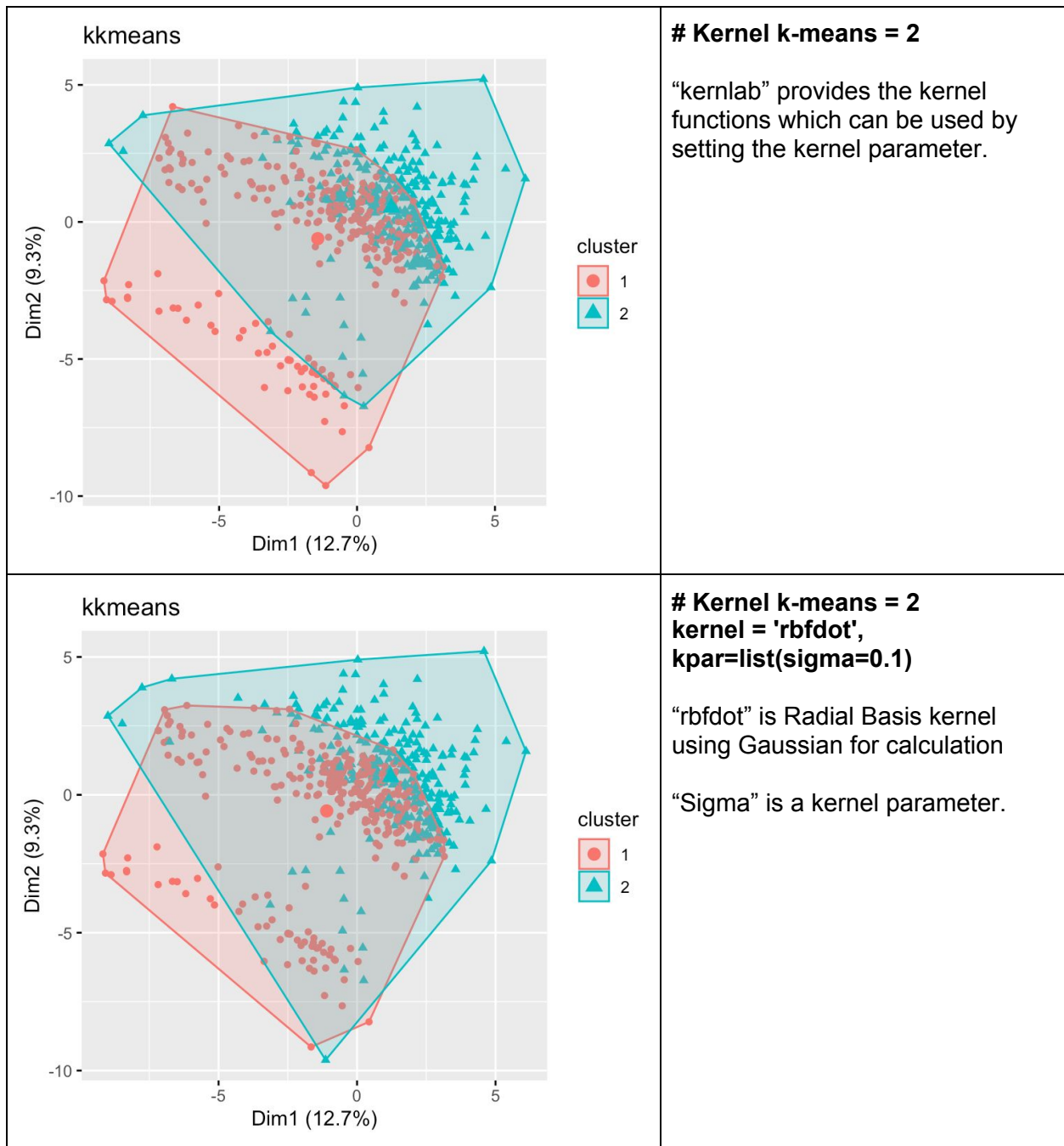
We tried different Parameters for DBSCAN clustering  
Results show in the table

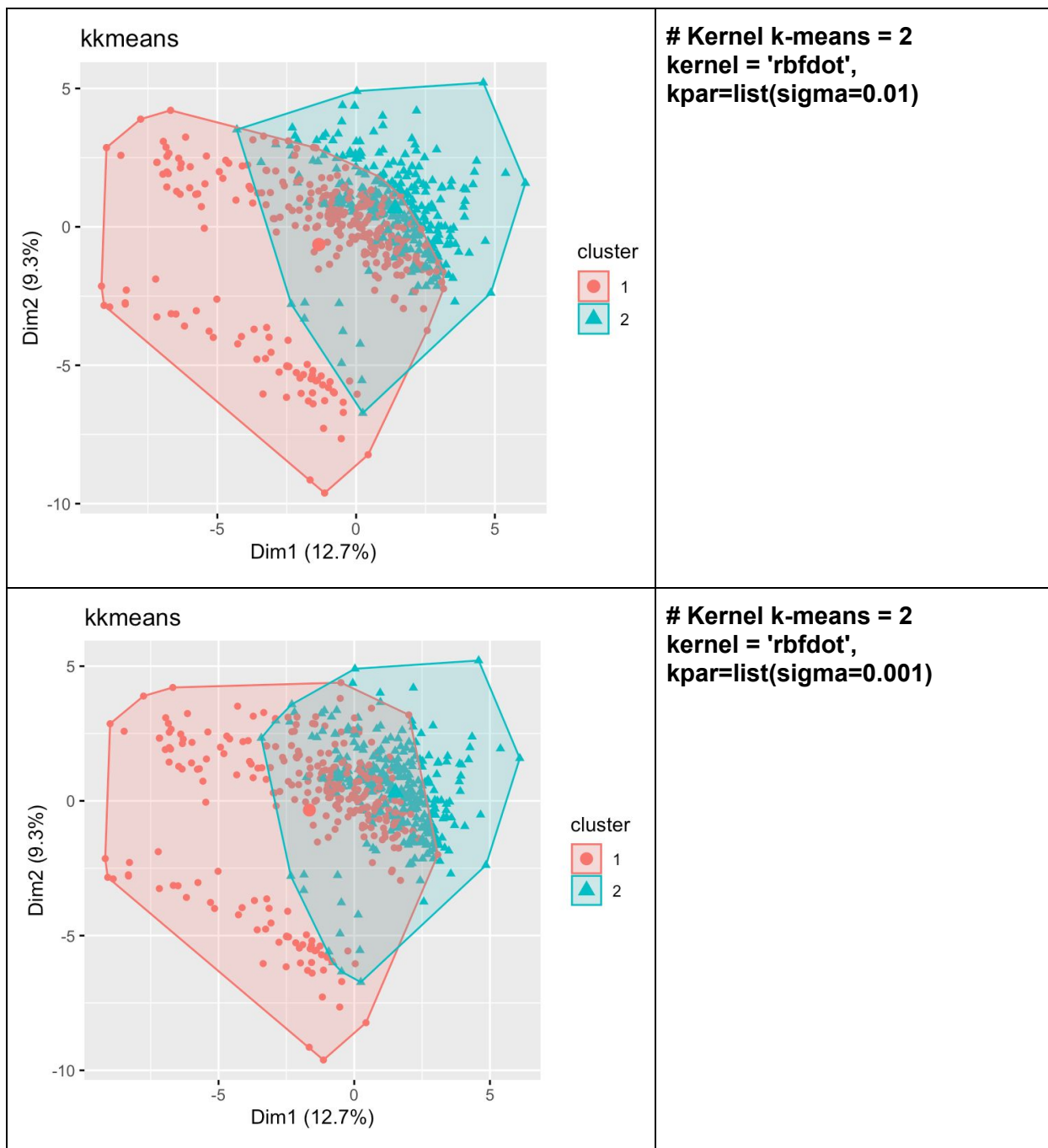
<p>dbscan</p> 	<pre>msDbscan&lt;-dbscan(bsd[,1:26], eps = 0.5, minPts= 10)</pre> <p>Black points indicates all outliers We tried to adjust the parameters in DBSCAN. We think that our data contains so many outliers. Normally, DBSCAN should treat well with data with outliers; however, it does not work because each data point is too far away from others. Thus, it is treated to be all outliers.</p>
<p>Cluster plot</p> 	<pre>msDbscan&lt;-dbscan(x, eps = 0.1, minPts= 10)</pre>

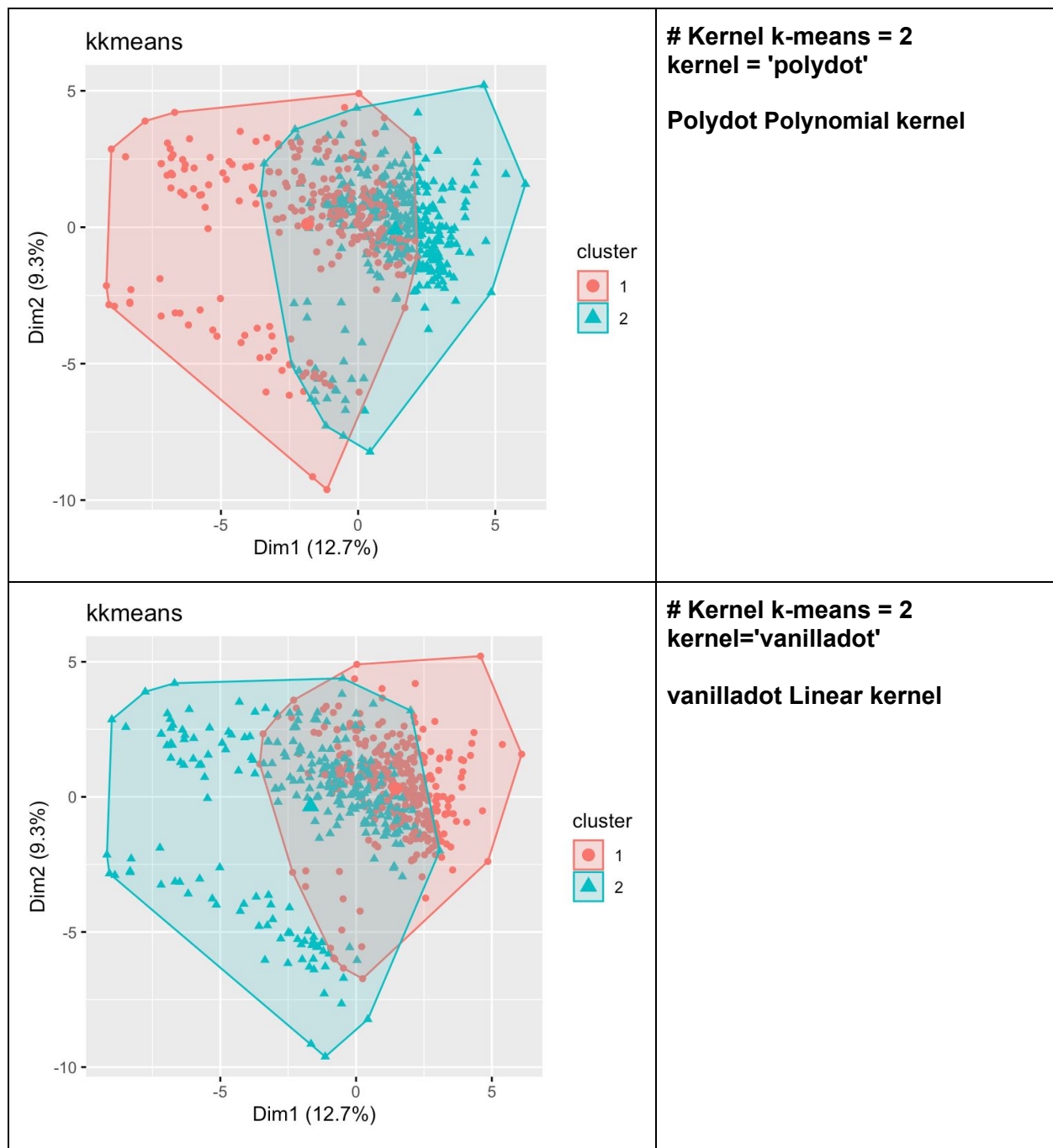


## Kernel-k-means

Trying Kernel KMeans with different values of  $k = 2$  with different parameters.







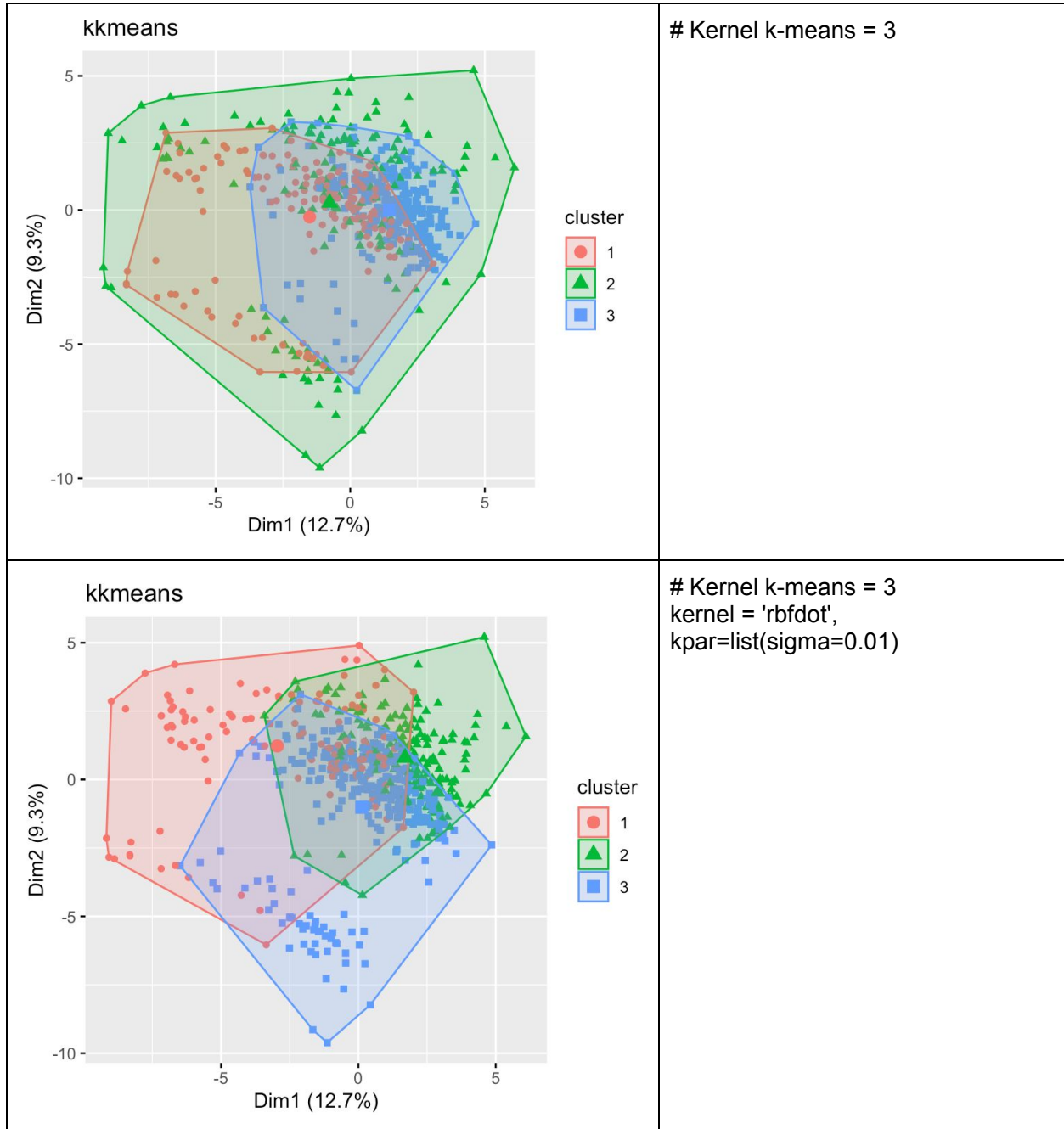
The prediction from k-means algorithm works not well on this dataset since it does not fit perfectly on generating clusters.

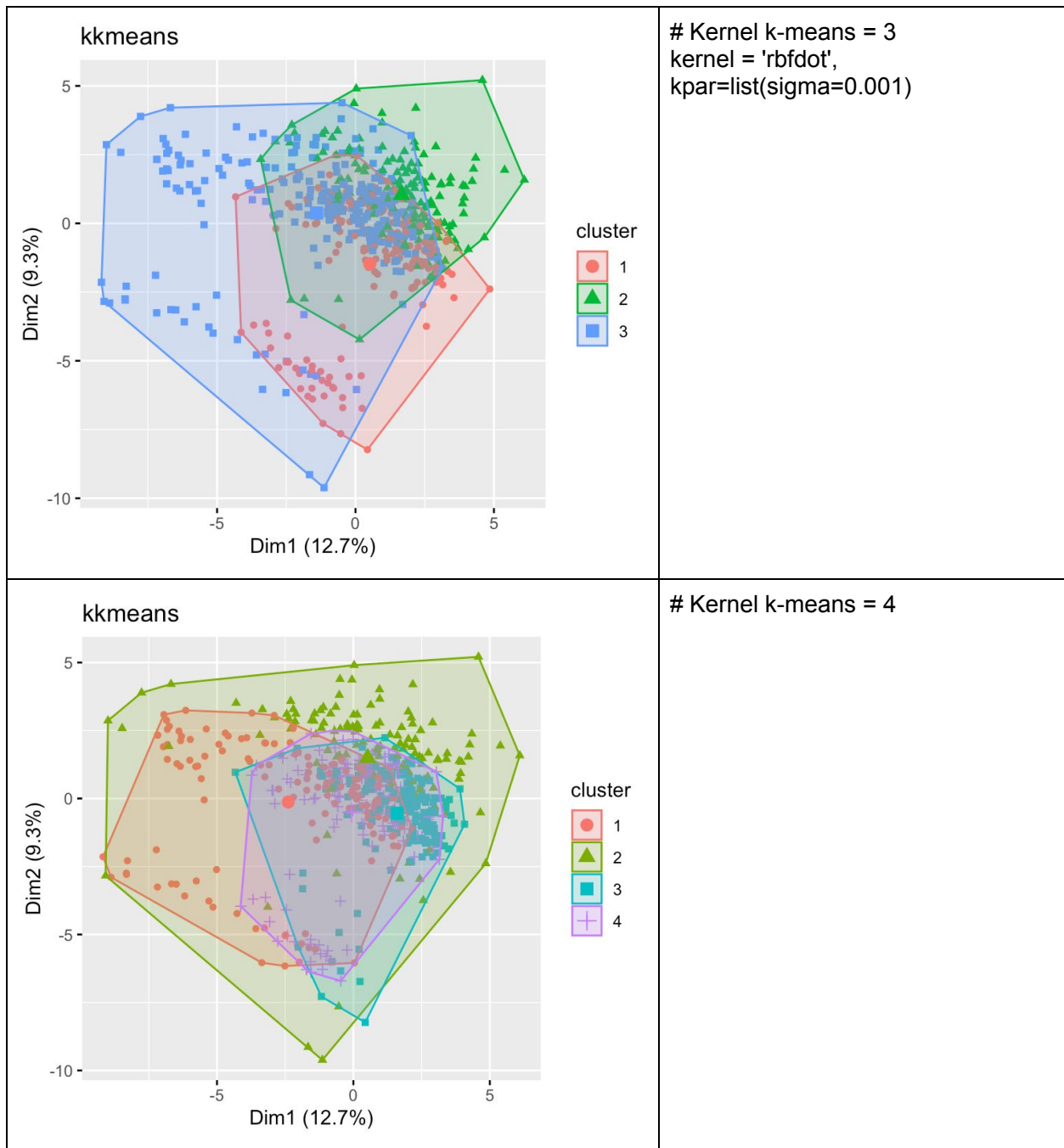
We tried different hyperparameters on the kkmeans method in the kernlab library.

Sigma is a radial basis kernel function. When we decrease the number of sigma, the overlapped regions are less down.

We also tried different kernels; vanilladot (linear kernel) and polydot (Polynomial kernel) which are the different mathematical methods for the kernel calculation. We got almost similar results for clustering.

Trying Kernel KMeans with different values of  $k = 3$  and  $4$  with different parameters.





We changed the number of Kernel k-means. The overlapped regions still presented. The best clustering by our basic observation is Kernel k-means equals 3 with rbfdot and sigma of 0.01.

**5. (a) Compare the clusters obtained in Q3 and Q4. Are the clusters obtained from the different procedures similar/different? Describe how they are similar/different – in terms of number and size of clusters, within cluster spread and separation between clusters; also, very importantly, interpretability.**

From the previous question 3 we perform clustering using the “kmean” and question 4 we have done the “agglomerative clustering”, “kernel-kmeans”, and “DBSCAN”

Clusters obtained from different methods bring about the different results.

Kmean gives the best result for clustering in this problem.

Agglomerative clustering results depends on the cutting point that we selected

Kernel-k-means is similar to Kmean but the overlaps area is larger than Kmean. Therefore, it is not suitable for selection.

DBSCAN doesn't work for this case. Our DBSCAN model indicates all data points as outliers.

All these methods are formulated differently and suitable for different type of datasets.

#### K-Means Cluster

Advantages of K-Means

- Easy to understand and implement.

Disadvantages of K-Means

- Sensitive to the number of clusters.  
After using techniques like Elbow method, it is sometimes hard to generate good clusters.
- Does not work well with outliers. Centroids can get dragged by the outliers resulting in skewed clusters.

#### Kernel- Kmean

Advantages of Kernel Kmean

- Kernel-based clustering algorithms can capture the non-linear structure in data.
- Suitable for clustered exceeds thousands clusters per dataset.
- Dramatically reduced run-time complexity

#### Hierarchical Clustering

Advantages of Hierarchical Clustering

- Easier to decide on the number of clusters by looking at the dendrogram.
- Easy to implement

Disadvantages of Hierarchical Clustering

- Difficult to predict the number of clusters (K-Value)
- Not suitable for large datasets
- Very sensitive to outliers

#### DBSCAN

Advantages of DBSCAN

- Work well for noisy datasets.
- Identity Outliers easily.
- Clusters can take any irregular shape unlike K-Means where clusters are more or less spherical.

#### Disadvantages of DBSCAN

- Does not work very well for sparse datasets or datasets with varying density.
- Sensitive to eps and minPts parameters.

**(b) Select what you think is the 'best' segmentation - explain why you think this is the 'best'. You can also decide on multiple segmentations, based on different criteria -- for example, based on purchase behavior, or basis for purchase,....(think about how different clusters may be useful.**

The best model from all of the models is K means clustering with  $k = 4$  because it gives us the sensible number of clusters and size of clusters as well. The spread within cluster spread and separation between clusters are better than other clustering methods because of less area of overlap. Different numbers of clusters result in different final characteristics of each cluster.

From the results we got in our data analysis and optimization we can conclude that the K means model with the 4 clusters is the best model. Within this analysis we came to a conclusion that we think fits CRISA market demand needs for advertising and marketing based on purchase behavior and basis for purchase. You can tell through the research above that there are many components that go into what could be the best market segmentation to promote to. Knowing your data and where the information comes from is crucial in understanding exactly what the task is at hand. Through our analysis we covered a breakdown and business goal of our clustering model along with descriptive information about household goods and demographic information in what we concluded was the area in western India that the survey took place.

	clusKM	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No_of_Brands	No_of_Trans	Brand_Runs	Total_Volume	Value	Trans_Brand_Runs
1	1	2.70	4.71	1.65	3.52	13.89	3.11	0.72	2.96	22.57	8.19	10280.09	1003.58	3.97
2	2	2.32	4.73	1.90	4.83	21.30	3.26	0.23	5.16	46.73	26.51	13197.14	1604.91	1.85
3	3	2.83	6.70	2.00	4.20	18.80	3.48	0.40	3.54	36.87	16.02	29981.09	3186.11	2.84
4	4	2.41	4.37	1.60	3.71	15.24	3.19	0.18	2.78	22.31	12.03	7799.24	934.67	2.08

In addition, we can create marketing campaigns from the result of clustering. For example, if we're looking on Cluster 2 which has the highest affluent index. This can be implied that this group of customers has the ability to pay. Furthermore, their brand loyalty is very low compared to other groups. And we know the demographic of these customers that they are mostly women with the specific age. We can make the marketing directly to this specific group in order to capture these potential customers and make them have more loyal to the brand that we promote. This is one way to make a permanent customer base for the company.

Based on everything we concluded with our market analysis for CRISA to make more cost-effective promotions in the appropriate segments for brand loyalty and customer rewards programs. From the table above (top performing), CRISA should focus on the transaction of

brand runs and the underlying methodology behind that variable. To focus on the transaction of brand runs we would provide them with the information that we also received directly responding from the number of transactions divided by the brand runs. Showing CRISA that if they can segment the market of customer needs by the total number of runs of purchasing the same brand by the transactions this would give them the top market for their advertising and marketing campaign.



**(c) For one 'best' segmentation, obtain a description of the clusters by building a decision tree to help describe the clusters. How effective is the tree in helping explaining/interpreting the cluster(s)? (explain why/why not). Does the decision tree provide a similar interpretation to that you find from the description of cluster centers; does it provide alternate or additional information which will be useful in understanding the clusters. (Note - you may develop decision trees for alternate clustering, and use these to help choose the 'best' clustering).**

For our clustering by building a decision tree we went with a hierarchical clustering model for our visual representation. Compared to the other methods in describing the relationship in our data we do not find decision trees as useful as other methods for analysis. Decision trees lack in depth information that is built into the model. Though the visual representation is impressive in what it is showing with the model, all the information through all the nodes and leafs is lost or hidden, whereas clustering gives a direct correlation with information directly embedded in the visual representation. You can see the differences in clustering between the clusters and the information within them.

