Humboldt University Berlin

Institute of Marketing

Prof. Dr. Daniel Klapper & Dr. Narine Yegoryan

Customer Analytics and Customer Insights

WS 2020/21

Special Work Performance 3: Describing Data, Segmentation & Clustering

Your answers including all tables and graphs must not exceed 8 pages (no appendix). Please start a new page when providing your report to a new subtask. Please use typeface Times Roman in 12pt with 1.15 line spacing (in tables and graphs you may use 10pt and 1.0 line spacing) and 1 inch space on all sides. Do not forget to report your names, group number, and student numbers and a page number on each page starting with number one on the first answering page. Do not include a title page or content page.
Send your team report as pdf to my email address daniel.klapper@hu-berlin.de not later than Jan 22, 2021, 4:00pm.

Download the following data sets from Moodle: README.txt, indivData.csv.

SWP 3a:
Get familiar with the data from indivData.csv. A description of this data is provided in the README.txt document.
Describe the basic structure of the data and report interesting findings. This documentation should explain what you did and why. The key findings should be reported in tables and graphs and they must also be discussed in the text (maximum 5 pages).

SWP 3b:
Use the data from indivData.csv and find meaningful segments/cluster among the 593 respondents. Describe your segmentation/clustering approach in a detailed way and also describe the clusters. Use tables and graphs so additionally interpret the clusters (maximum 5 pages).

SWP 3 – Describing Data, Segmentation & Clustering

3a. Data Structure and Key Findings

The given dataset consists of the results derived from an exploratory market research questionnaire, regarding the respondent's personal information, as well as involvement, preferences and opinions regarding Bluetooth speakers. Some answers had reversed answer scales and patterns in order to check for respondent's attention and prevent unusable answer sets.
The resulting data basis indivData consists of 593 respondents answering 36 answers each. 3% of the participants did not indicate their gender and therefore will be dropped from the subsequent analysis data basis due to inconclusive answers. Furthermore, we will exclude the 13% of respondents that did not state their monthly income, due to the implications that this uncertainty holds for marketing research purposes. The cleaned data set consist of 506 entries. The two largest economic subsets consist of 57.90% students and 31.82% employed people. Roughly half of respondents, 57.90% stated a monthly household income below 1000€. 82.60% of respondents are aged between 18 and 29, indicating a rather young segment. 57.11% of respondents stated their country of residence as Germany, while the rest was widely spread across 30 different countries. Due to the nature of Bluetooth speakers, there are no special implications derived here, due to technical similarities across markets and the lack of local distribution. 54.55% of respondents are male, 45.45% are female.

The questionnaire comprised multiple Bluetooth speaker properties the participants had to value. They were asked to distribute 100% across the four attributes "Battery", "Price", "Sound" and "Weight". The respondents valued the importance of the attributes as shown in Figure 1.

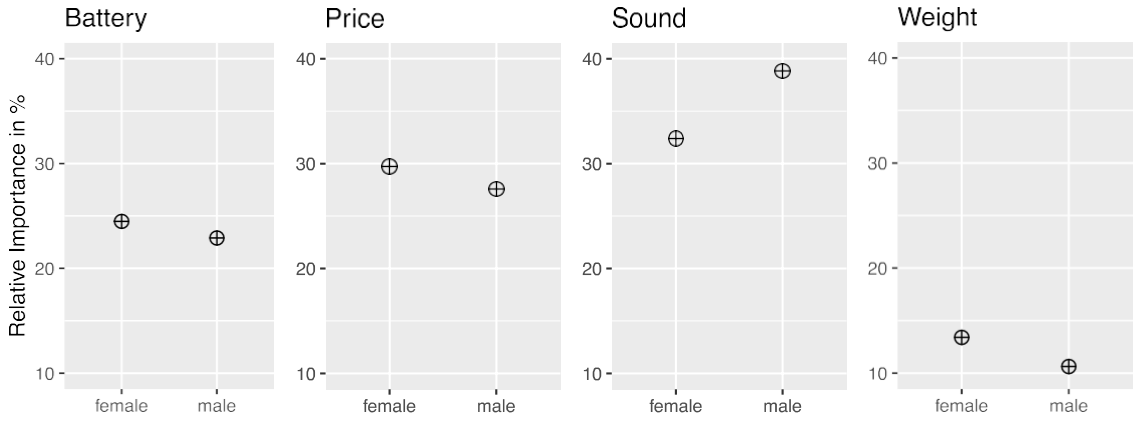

Figure 1. Relative Importance Rating of Properties (in %)

Multiple implications can be derived from that result. The sound is valued by the participants as the most important property of Bluetooth speakers, while males tend to value it slightly higher than females. The second most important factor is the price of the speaker. Since 57.9% of participants have an income lower than 1000€ this is not a surprising but an important finding. Figure 1 clearly shows where the focus of a marketer has to be, namely on the sound

and price. In combination with the fact that 82.6% of the respondents are aged between 18 and 29 years, we can state that when targeting the participants of the presented study, marketers or sellers should focus to present the Bluetooth speakers as a low-priced but still high-quality good, as well as presenting it as a young and cool device. A light weight scarcely increases the value of a Bluetooth speaker. When taking gender into account within the four properties, it can be stated that females are willing to forego sound quality in order to get a lower-priced product. Thus, from the marketer's point of view it can be advantageous to differentiate the product in that regard. When differentiating the four properties shown in Figure 1 by the two most important age groups (18-24 years and 25-29 years) we can state that with increasing age, the preferences slightly shift. The respondents above an age of 24 years valued the sound and the battery capacity slightly higher than the 18–24-year-old group. For that increase in product quality, they accept a corresponding increase in price. A discrimination in terms of this finding might be considered when selling Bluetooth speakers to the considered age groups. The aforementioned implications are derived by the results shown in Figure 2. The presented numbers are means across the age groups in percent.

Age	**Battery**	**Price**	**Sound**	**Weight**
18-24 years	23.50	29.59	35.13	11.77
25-29 years	24.39	27.62	36.60	11.40

Figure 2. Relative Importance Rating in % (by age group)

In marketing, knowing your potential consumer is essential if not even the most important aspect. The more detail a marketer or business has about the consumers, the more specific can the groups be targeted, which ideally results in higher sales. With respect to understanding the consumers of Bluetooth speakers in the given data set, we decided to first filter the data by people who already own a Bluetooth speaker. In the given data set, 234 out of 506 participants already own a Bluetooth speaker, that corresponds to a share of 46.25%. From these 234 people, 191 are between 18 and 29 years old, consisting of 79 females (41.36%) and 112 males (58.64%). This suggests that the main consumers are young men between 18 and 29 years. The consequences of defining this demographic group as the main consumer would be as follows when we take into account the results from property analysis. The sound of the Bluetooth speaker would be the most important attribute. Furthermore, the consumer group would be willing pay a higher price when they perceive the sound as being high-quality. In the context of consumer targeting, it is even more important to take a deeper look at the variable "IntentToBuy" because it may occur that that group varies from the group already owning the good. In the given data set 174 respondents intent to buy a Bluetooth speaker. 151 of them are in the age between 18 and 29 years, of which 45.70% are female and 54.30% are male. It can be stated that this result regarding the age distribution corresponds to the findings about participants who already own a Bluetooth speaker.

Through further differentiation of the group of 151 people who intend to buy a Bluetooth speaker we were able to derive the following results. Among the people who already own a speaker, 51 intend to buy a new one. From these 51 people 19 people are female and 32 people are male. This is not a surprising result when considering the previous findings. As stated before, males value the sound of the speakers as the most important attribute. With technical

progress the sound quality gets better with every new generation of Bluetooth speaker. Thus, males buy Bluetooth speaker more frequently in order to get better sound quality. Females in the given data set, however, are more price sensitive in the context of Bluetooth speakers and do not value the sound as important as males do. Therefore, they may use a Bluetooth speaker over a longer period of time. An interesting finding is that among the people in the data set who do not own a Bluetooth speaker but intend to buy one the share of females and males is 50:50. That implies, that when looking at consumers between 18 and 29 years who do not own a Bluetooth speaker, males and females should be targeted in the same proportion, although young males previously seemed to be the main consumer group. When creating follow up marketing campaigns, however, it might be more efficient to target males rather than females, as shown above.

In order to further specify the marketing strategy, we decided to group the participants aged between 18 and 29 years who intend to buy a Bluetooth speaker based on their subjective knowledge regarding the product. The variable "Subjective Knowledge" is built by the mean of 5 different subjective knowledge measures in the questionnaire, gathered for every participant. It can take values from 1 to 7. Group 1 is defined by respondents who stated their subjective knowledge between 1 and 3 (44 people), Group 2 consists of the people who stated their subjective knowledge between 3 and 5 (71 people) and Group 3 contains every person who stated its subjective knowledge above 5 (36 people). This split across leads to the following results (Figure 3).

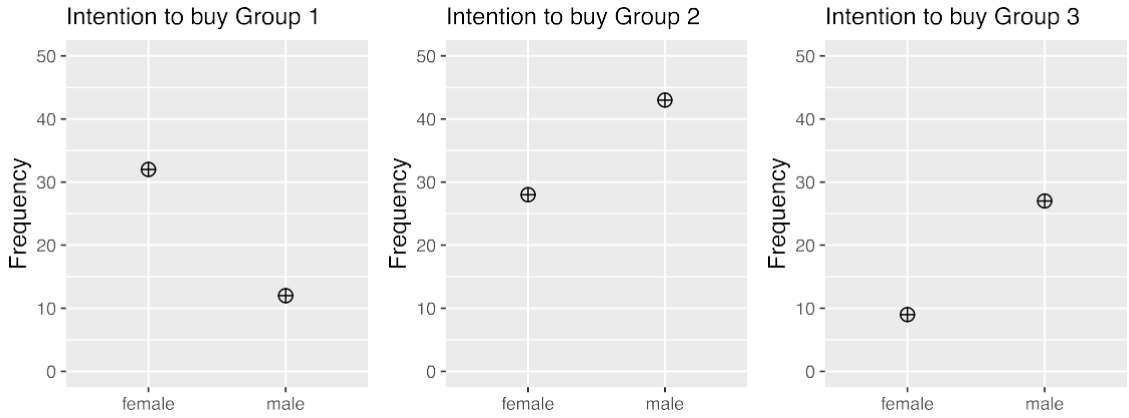

Figure 3. Intention to buy among Subjective Knowledge Groups (age 18-29)

Most of the participants who intent to buy a Bluetooth speaker have a subjective product knowledge between 3 and 5 (Group 2). This is neither very high nor very low, they have a rather average subjective knowledge about the product. In Group 2 (medium subjective knowledge) as well as in Group 3 (high subjective knowledge), males are overrepresented. The females in the given data set rather perceive their product knowledge as low, therefore most of them show up in Group 1. These findings are consistent with the previous analysis. As already stated, males have a higher focus on the sound of Bluetooth speakers than females do (Figure 1). That can now be explained by Figure 2. Since males perceive their product knowledge as higher than females do, it consequently influences their buying decision in that direction. The same effect holds for females, who base their buying decision mainly on the product price.

The presented findings suggest that males might have a higher involvement in the product than females do. The higher the consumer's product involvement, the higher the probability they are going to buy the product. In order to evaluate the participant's involvement levels, we built groups in the same manner we did with the subjective knowledge (age 18-29, "IntentToBuy" = 1). Again, most respondent's contained in Group 2 with a medium product involvement between 3 and 5. Marketers should therefore prepare for that involvement level when creating advertisements for Bluetooth speakers. Another challenge for the businesses is to increase the product involvement in the long term in order to increase sales. In contrast to the subjective knowledge, there are more males than females in every Group.

Group 1		**Group 2**		**Group 3**	
Males	Females	Males	Females	Males	Females
10	7	62	57	10	5

Figure 4. Personal Product Involvement Groups (age 18-29)

When taking the results from the analysis about who does not own a Bluetooth speaker but intents to buy one into account, we can conclude that one task of businesses is to increase the product involvement of females. They are as important as male consumers, with the difference that they have a lower perceived subjective knowledge (Figure 3) and therefore may focus on slightly other things than males (Figure 2). When increasing the product involvement, subjective knowledge might increase and that can lead to more customers. Therefore, in the long term it is important to increase the female as well as the male consumers product involvement to maximize sales.

3b. Cluster Analysis

In the first step three columns were added to the dataset, each containing the mean of the brand awareness, subjective knowledge and involvement values, as they would be useful for the cluster analysis and the evaluation and interpretation of the clusters.
For the clustering it was necessary to find a subset of the data that contained a sufficient amount of meaningful variables to yield good clusters but didn't contain too many variables, so the clusters would be sufficiently spaced for the algorithm to correctly identify them. In the graph below one can see the MDS of three different subsets.

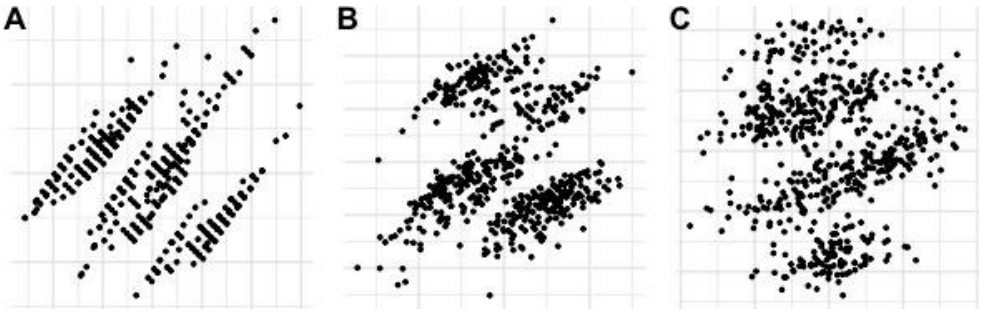

Figure 5. Multidimensional Scaling of various subsets

Because there were no clusters visible in the MDS of the entire dataset it was necessary to reduce the amount of variables used. We tried out different combinations of variables and it quickly became apparent that the variables Own and IntentToBuy would be the most relevant for the cluster analysis, because without them no distinct clusters would be visible. We therefore started with those two and gradually added additional variables. In the figure above the subsets with the most promising combinations of variables are displayed.
A is the MDS of a subset containing the variables Own, IntentToBuy, RelImp_Price and Income. Three distinct and well spaced clusters are visible. The subset used for B contains the same variables as A but additionally the above mentioned average of knowledge. In subset C the averages of brand awareness and involvement are included in addition to the variables of subset B. The MDS of subset B and C show four clusters, which is promising. Unfortunately subset B yielded the same results as subset A, except that DBSCAN produced significantly more outliers. The clusters in subset C on the other hand are not sufficiently distinct for any algorithm to achieve a satisfying result. Therefore we settled on using the subset whose MDS is displayed in graph A for the analysis.
In the next step we tried out several different clustering algorithms to find the one which achieves the best result. The graphical results of the cluster analysis can be seen in the figure on the next page. While Hierarchical Clustering, K-Means Clustering and Gaussian Mixture Modelling (GMM) have achieved decent results, only Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was able to correctly identify the three clusters. This makes sense as the clusters are well spaced and have roughly the same density. A great thing about DBSCAN is also that it marks points that lie alone in low-density regions (i.e. whose nearest neighbors are outside of the set epsilon neighborhood) as outliers and adds them to another cluster.

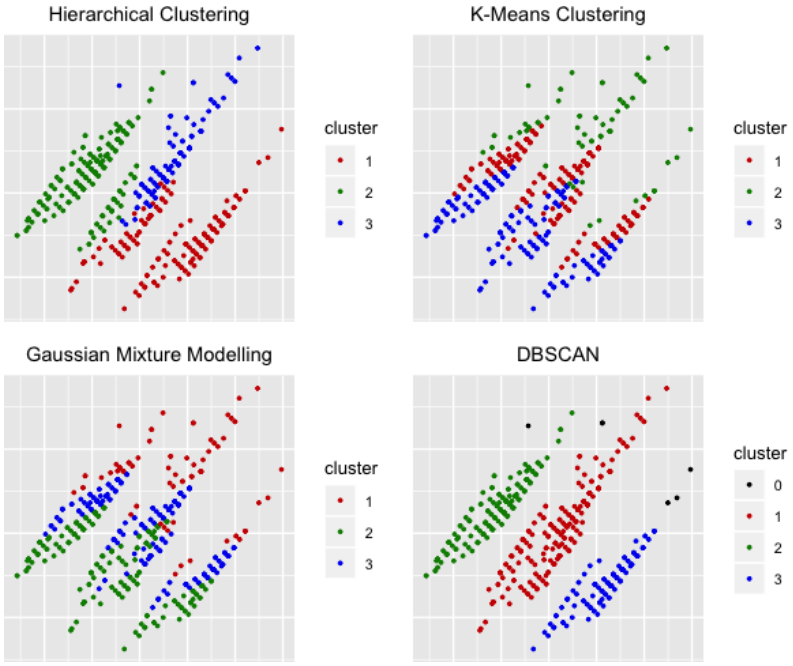

Figure 6. Graphical results of the cluster analysis

Below one can see the mean values for each cluster created by DBSCAN and each variable, including the averages of brand awareness, knowledge and involvement at the end.

	Group.1	id	Own	IntentToBuy	BrandAwareness_Anker	BrandAwareness_Bose	BrandAwareness_JBL	BrandAwareness_Philips
1	0	621.6667	0.1666667	0.5000000	0.16666667	0.6666667	0.5000000	0.3333333
2	1	926.1992	0.2549801	0.2549801	0.09163347	0.5936255	0.4980080	0.3665339
3	2	966.3171	1.0000000	0.0000000	0.17073171	0.7219512	0.6878049	0.3902439
4	3	907.7252	0.0000000	1.0000000	0.11450382	0.6870229	0.6488550	0.4122137

	BrandAwareness_Sony	BrandAwareness_UE	BrandAwareness_HarmanKardon	BrandAwareness_Beats	SubjKnow_r1	SubjKnow_r2	SubjKnow_r3
1	0.5000000	0.16666667	0.16666667	0.1666667	2.666667	2.833333	2.333333
2	0.5418327	0.09561753	0.09561753	0.4302789	3.318725	3.577689	2.470120
3	0.5365854	0.22439024	0.18048780	0.4975610	4.302439	4.478049	3.439024
4	0.5954198	0.09923664	0.12213740	0.5725191	3.679389	4.015267	2.862595

	SubjKnow_r4	SubjKnow_r5	PII_1	PII_2	PII_3	PII_4	PII_5	RelImp_battery	RelImp_weight	RelImp_price	RelImp_sound
1	4.000000	3.166667	4.333333	4.500000	4.333333	4.166667	3.833333	6.333333	1.00000	86.66667	6.00000
2	4.135458	3.557769	3.705179	3.780876	3.812749	3.737052	3.689243	23.478088	11.92829	30.49402	34.09960
3	4.741463	4.746341	4.770732	4.858537	4.887805	4.785366	4.541463	23.795122	10.53659	26.34634	39.32195
4	4.389313	4.106870	4.503817	4.595420	4.603053	4.496183	4.473282	24.312977	13.12214	25.88550	36.67939

	Gender	Age	Occupation	Education	Income	brand_awareness	knowledge	involvement
1	1.666667	3.333333	2.000000	3.000000	4.833333	0.3333333	3.000000	4.233333
2	1.565737	3.111554	2.318725	3.187251	3.418327	0.3391434	3.411952	3.745020
3	1.643902	3.180488	2.292683	3.239024	3.736585	0.4262195	4.341463	4.768780
4	1.564885	2.931298	2.305344	3.290076	3.297710	0.4064885	3.810687	4.534351

Figure 7. Numeric results of the DBSCAN

Cluster 0 consists of six outliers. All of them seem to have rather low subjective knowledge about Bluetooth speakers. However, Cluster 0 is not significant since it only contains six elements. Therefore we excluded it from further analysis. Cluster 1 are the “Scrimpers” with 251 elements. They mostly don’t own and don’t want to buy a bluetooth speaker. Overall they are quite average, but have little knowledge and generally care about the price. The second cluster contains the “Owners” with 205 elements. They all own and don’t want to buy a speaker, have high knowledge, high involvement and care about sound. Furthermore they have the highest age and the highest income out of the three main clusters. Finally, cluster 3 are the “Poor Graduates” with 131 elements. They all don’t own a bluetooth speaker, but want to buy one and show rather high involvement. They are the youngest and have the lowest income, but also the highest education level.

Cluster 1

The "Scrimpers" are the most average of all segments here, but there are several important deductions to be made, to better understand their motives, and possible solutions or strategies to approach them. From a purely economic standpoint they are an attractive group, overall average income, but the largest of any of the groups, thus making up a large part of the respondents overall buying power. In terms of attractiveness for a Bluetooth manufacturer to market to them is quite a multi-faceted issue. While the "Scrimpers" ranking lowest among all segments regarding overall brand awareness, involvement and knowledge, could be interpreted a general disinterest for the Bluetooth speaker market, there is also a potential upside to that. Low involvement and product knowledge generally lead to lower demands regarding technical capabilities. Pairing that with the high relative ranking of price importance, cluster 1 seems like an ideal target group for a low-end, low-priced product category. And while it might seem difficult to convince a potential customer segment to purchase a product that they do not care much about, their low rankings could also be interpreted as lack of information. Lack of information then in turn, coupled with low brand recognition makes the "Scrimpers" once again an ideal target segment, for brand recognition strategies, coupled with informationally loaded campaigns, that position the marketer's brand as the Status Quo.

Cluster 2

In Marketing Theory, researchers often refer to social reference groups in order to primarily and secondarily determine what social aspiration certain groups of people connect to specific products. Due to the abundant advertising that Beats by Dre does with international music and sports personalities, students mentally form a connection between the lifestyle that they aspire to and the product that is being advertised. Seeing a specific type of successful sports personality wear Beats repeatedly, will form a mental connection between the product and the lifestyle of those people, thus a social reference. These social references are also especially important when looking at Cluster 2, "The Owners". While those people do already own a Bluetooth speaker, their opinion and brand recognition are more important for marketers than the average customer is. Cluster 2 is very attractive as an audience for advertisements even despite their lack of intent to buy a new speaker, as they already own one. This is due to the relative importance that a possible changed opinion holds. In social groups there are always opinion drivers and opinion adaptors. By asking the respondents how their expertise regarding Bluetooth speakers rank among their peer group in their perception, those opinion drivers can be identified and are most present amongst Cluster 2. Making an impact on their opinion leads to a multifold increase in marketing efficiency due to the rippling impact that their personal choices and beliefs have, strengthened by their inherent psychological confirmation bias thus forming a feedback loop positively connected to increased brand awareness and positive association.

Cluster 3

Generally, marketing is very dependent upon how specific certain advertisements are to the target group, and how hard that specific target audience is to reach through the available advertising channels. Additionally, there is a second, demand driven layer that influences so called impression rates for advertisement spending, solely influenced by how much operational

value is expected to be derived from an advertisement. The second part especially makes advertising to more affluent target audiences more expensive for the most part, due to the higher disposable income available, leading to more price flexible advertisement options. Applying this basic principle in analysis to the aforementioned clustering leads to the discovery of a very strong opportunity to secure future high-income customer bases by strategically investing in advertisements now for a customer group that both wants to buy, has the future assets to do so, and ranks pricing importance rather low in favor of technical product metrics.

When further dissecting Cluster 3, the “Poor Graduates”, it makes them seem like a very hard to satisfy customer segment. Very low average income, thus little disposable income to spend on Bluetooth speakers, while still displaying persistently high demands regarding sound quality. Their demand profile makes them very unappealing, while their overall intent to purchase the product in turn makes them a contested customer group. While they may not currently have the disposable income to afford the most expensive products, there is a case to be made, that the expected, very strongly positive Compound Annual Growth Rate (CAGR) over the next 15 years for the average student, makes them very important potential future customers. With how cheaply that audience can be reached now, compared to 15 years in the future, overall efficiency is greatly increased, even including those that potentially drop out or do not become as successful. This means that the initially very unappealing-seeming customer segment of cluster 3, is especially valuable for premium-segment manufacturers.