Gym Attributes and Customer Preferences

Master in Data Science and Business Analytics – Bologna Business School
stats_amazing_team – 2023

Riccardo Mioli, Daniela Malavita, Beatrice Maria Vergati, Michelle Lee Kruschke

Table of Contents

Introduction:

Worldwide people are becoming increasingly more aware of the benefits of a healthy lifestyle, despite the fact that humans are also leading increasingly sedentary lives. The World Health Organization estimates that nearly 30% of the world is not active enough and states that inactivity negatively impacts healthcare systems, economic development, and community well-being. Because of this, working out and making time for fitness has become a growing trend for millions of people. A little under 200 million people across the globe use a gym. Furthermore, gyms and fitness clubs were a nearly 100 billion dollar industry as of 2020[1]. Although the pandemic may have temporarily interrupted some gym activity, according to the IHRSA, a global health and fitness association, the industry is expected to grow over 100% by 2028. Given that there is potentially a lot of opportunity in this sector, what is it that gym-goers look for when choosing a gym? This is precisely what this project endeavoured to discover.

The study chose to look at six distinct behavioural characteristics: gender, age, presence of children, employment status, car usage and workout frequency. In addition, eleven independent factors were selected to consider regarding attributes of the facility. The survey questioned 136 respondents on the importance of each aspect.

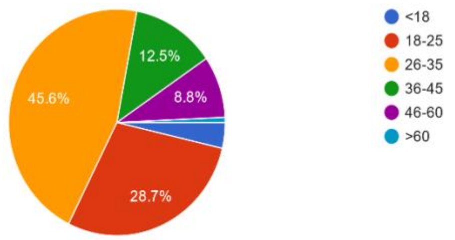
These endogenous questions related to price, location, workout preferences, amenities and operating hours. The survey was created in English with Google forms, distributed digitally and complied anonymously. Participants were asked to rank the importance or personal value of each of the gym attributes on a scale of 1-7, with 1 being of least importance and 7 being of the highest importance. Shown on pages 2-4, is a complete summary of the questions and responses to the survey that was used in the analysis. Subsequently all of the data was processed using SAS® analytics software in order to dive deeper into the results. The objective of the project was to develop a cluster analysis using a sequence of statistical methods.

[1]Taken from https://www.statista.com/topics/1141/health-and-fitness-clubs/#topicOverview

Survey description:

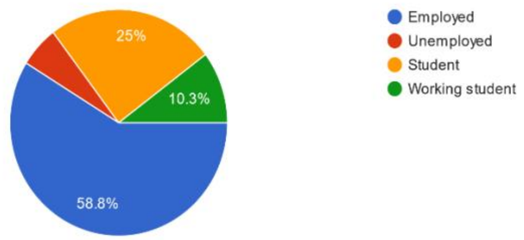
Select your age range.

136 responses



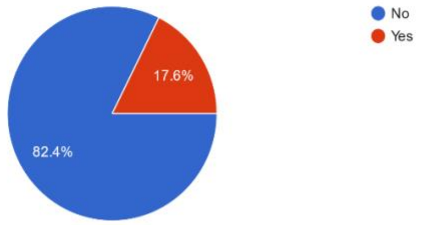
What is your current status?

136 responses



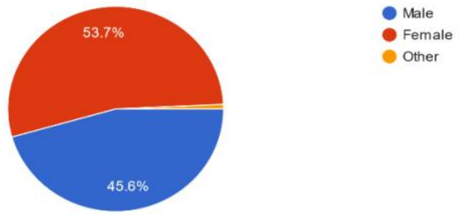
Do you have children?

136 responses



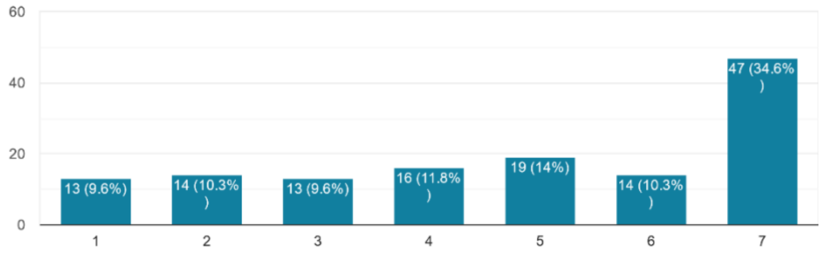
Select your gender.

136 responses



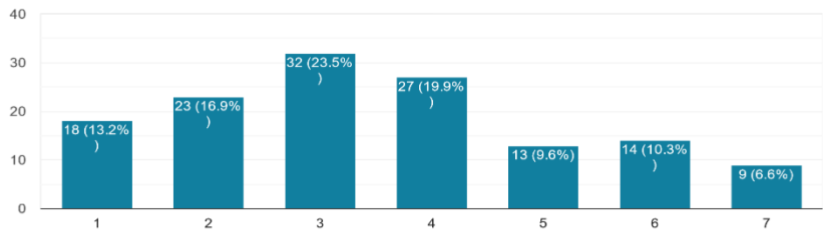
How often do you use a car to go to places?

136 responses



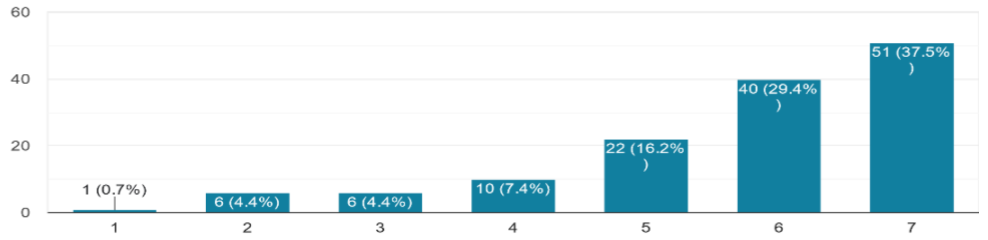
How often do you workout?

136 responses



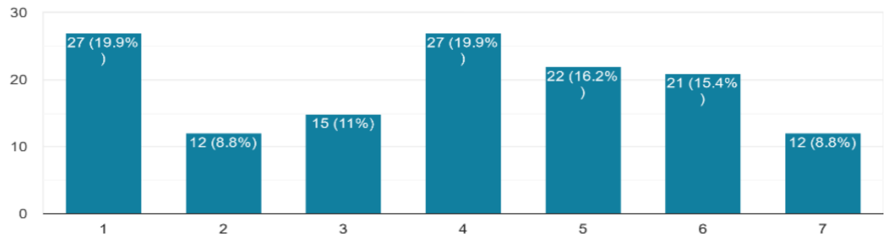
Price is an important factor when I choose a gym.

136 responses



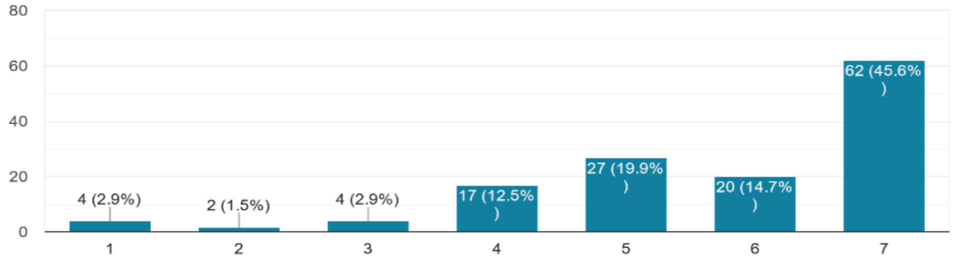
The possibility of paying an annual subscription in installments is important to me.

136 responses



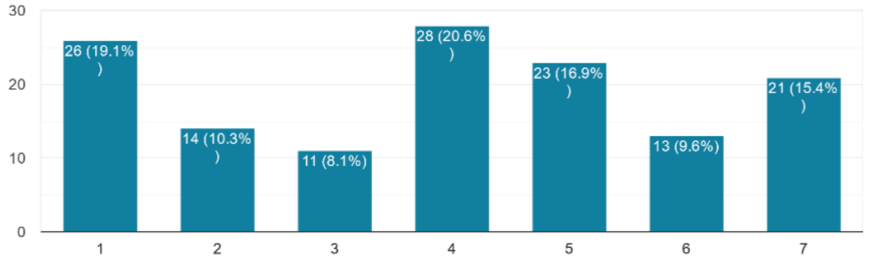
It is important to me for the gym to be close to my house.

136 responses



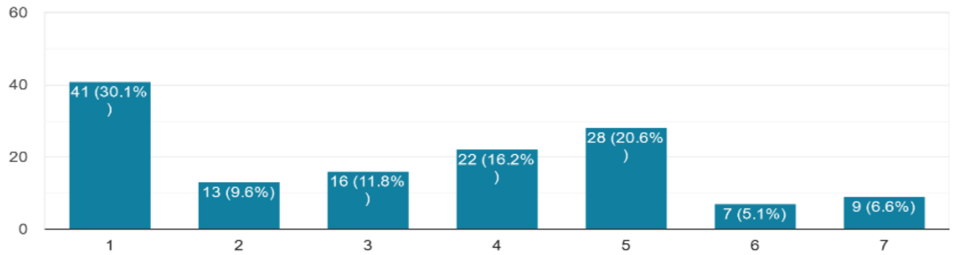
It is important to me for the gym to be close to my workplace.

136 responses



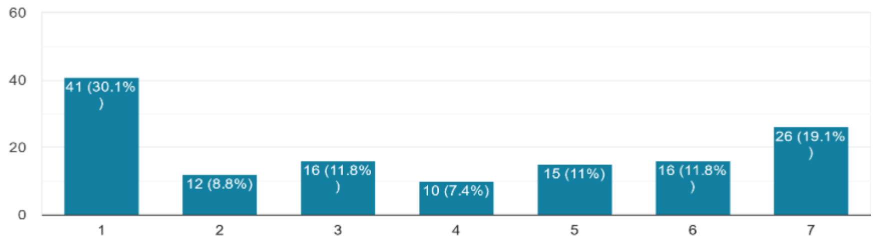
Training with friends is important to me.

136 responses



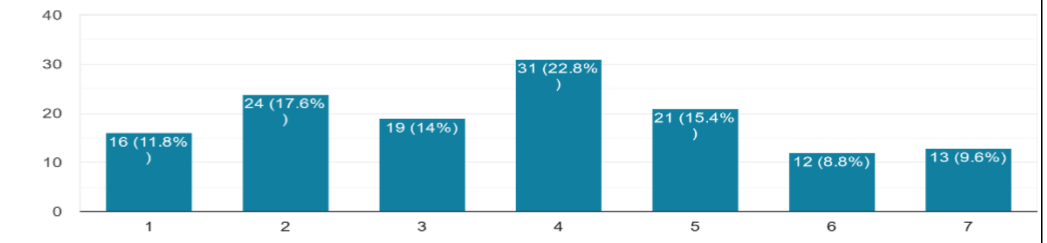
The presence of courses (e.g. yoga, total body, group cycling, etc.) is an important factor when it comes to choosing a gym.

136 responses



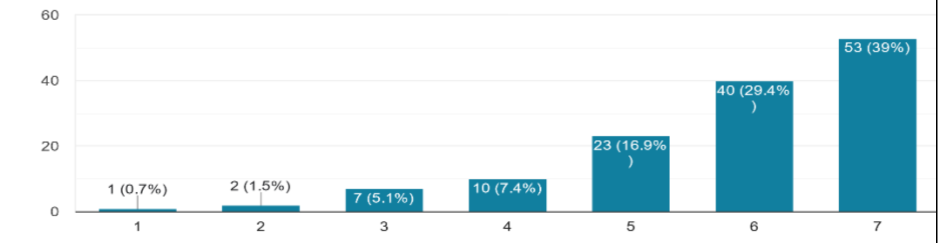
When I choose a gym, the presence of the newest training machines and their brand is important to me.

136 responses



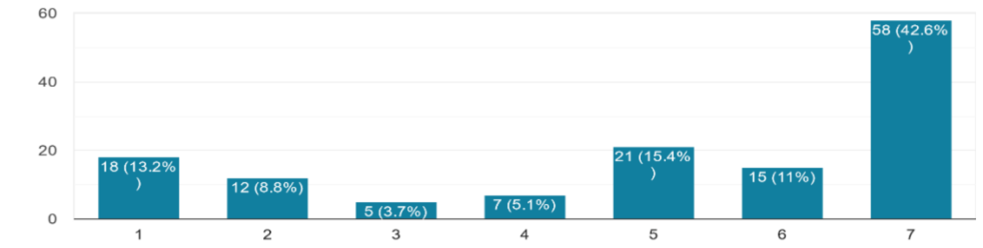
Great flexibility in operating hours (e.g. early in the morning/till midnight/open on weekends, etc.) plays an important role when I have to choose a gym.

136 responses



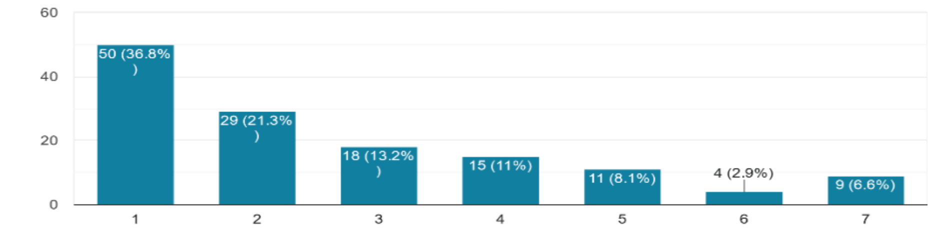
It is important for the gym to have available parking.

136 responses



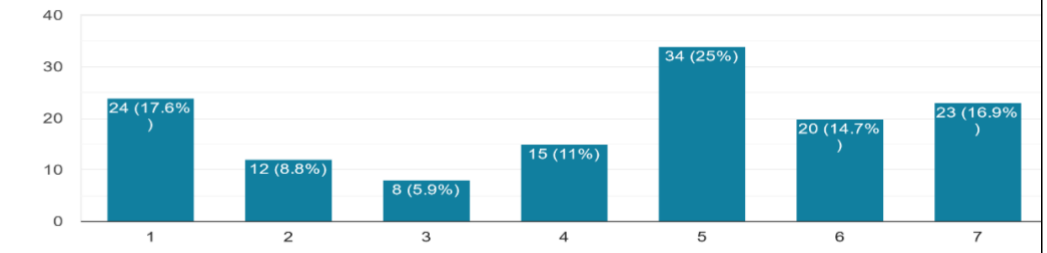
Access to amenities, such as saunas/massages/protein shops, is crucial to me when choosing a gym.

136 responses



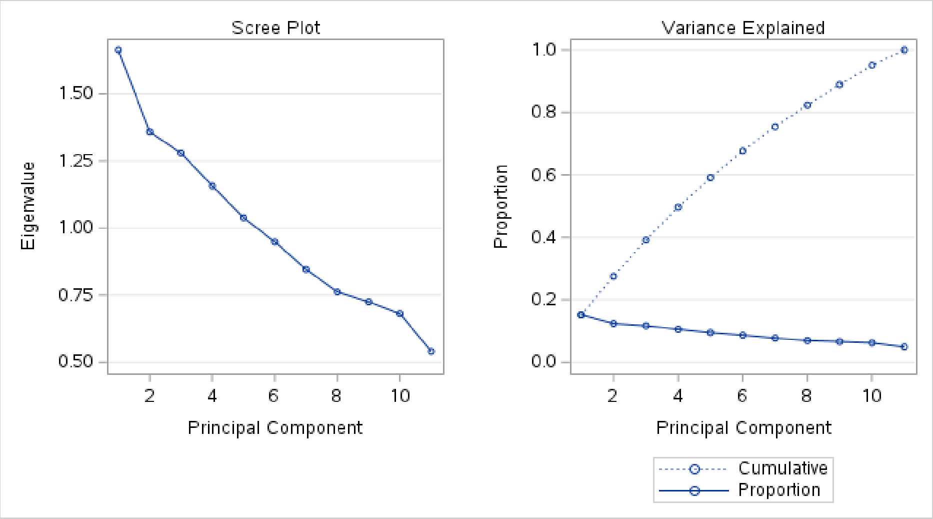
The possibility to get personalized programs from qualified personnel like a certified personal trainer, or a certified nutritionist, or from both is important to me.

136 responses



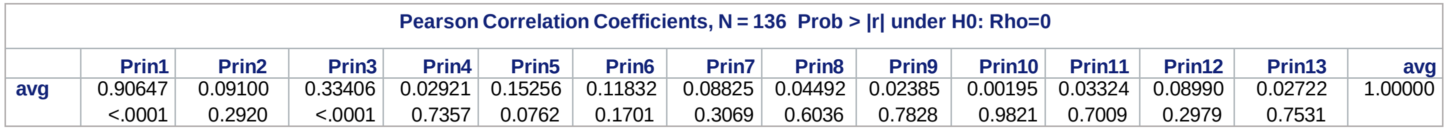
Analysis:

PCA and Size effect



The analysis began by simply looking at the mean and standard deviation for each of the scaled variables, including car usage and workout frequency. Next in the process a Principal Component Analysis (PCA) was run and the correlation among variable averages was checked, shown in the table below. The PCA values were used to evaluate the correlation between the average value of the individual variables and each principal component in order to verify the need to eliminate the size effect. As well, visible in the table below, principal component one shows a high correlation with the mean answer of each user. This means that "the respondents tend to provide their ratings by referring to a personal and latent mean vote".[2] To solve this problem, the data was scaled using a "non linear row standardization of the input matrix".[3] The scaling process is based on centering each user answer on the basis of the mean perceived by the user.[4]

After implementing the process for size effect removal an additional PCA was conducted over the scaled dataset. Six principal components were chosen using a threshold λ ~= 1 for eigenvalues which, shown in the graph on the left, explains the large 67.7% of the variance. The threshold for λ was chosen following the Kaiser heuristic, this way only the eigenvectors with a "sufficient" amount of variance are kept.



Pearson Correlation Coefficients, N = 136 Prob > |r| under H0: Rho=0

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13	avg
avg	0.90647 <.0001	0.09100 0.2920	0.33406 <.0001	0.02921 0.7357	0.15256 0.0762	0.11832 0.1701	0.08825 0.3069	0.04492 0.6036	0.02385 0.7828	0.00195 0.9821	0.03324 0.7009	0.08990 0.2979	0.02722 0.7531	1.00000

F. Camillo, V. Adorno. PLS-PM in opinion surveys when respondents minds generate size-effect (involvement axis) in the data[2]
Ibidem[3]
More information about the process can be found in F. Camillo, V. Adorno. PLS-PM in opinion...[4]

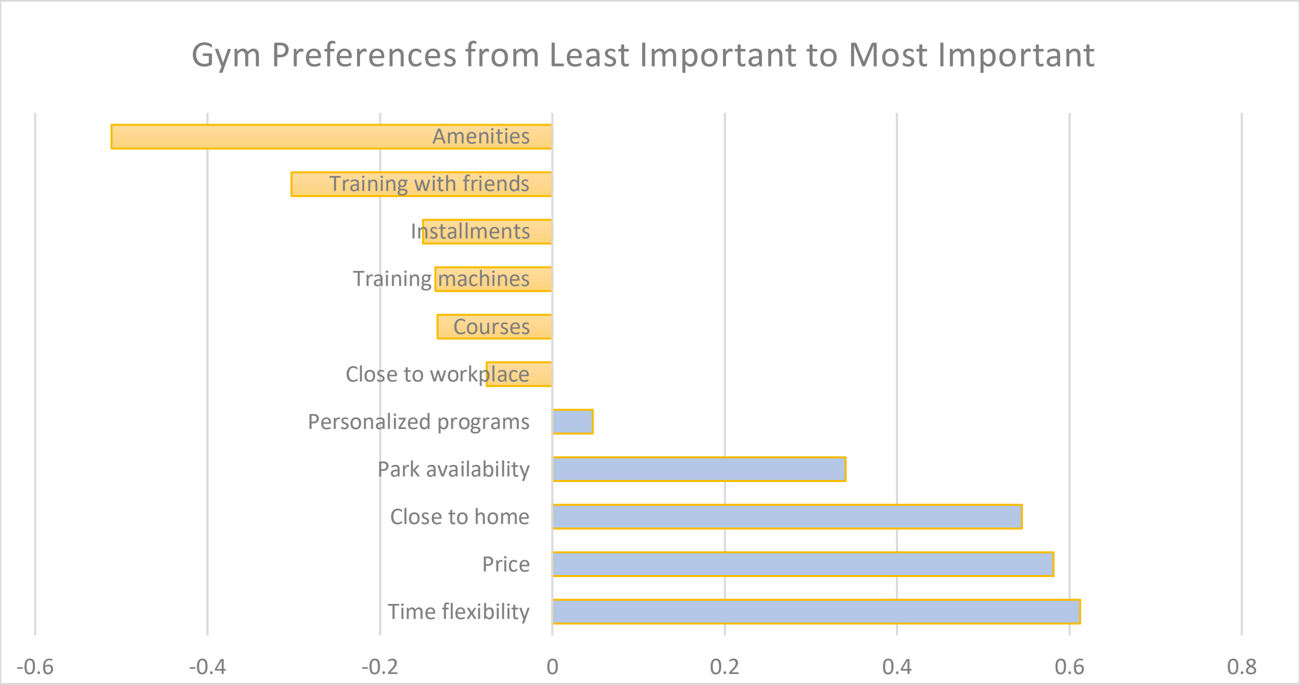
Attribute means after scaling:

We also looked at the mean and standard deviation of our 11 endogenous variables after size effect elimination.

nvar8	price
nvar9	installments
nvar10	closeness_home
nvar11	closeness_workplace
nvar12	training_wfriends
nvar13	courses
nvar14	training_machines
nvar15	time_flexibility
nvar16	park_availability
nvar17	amenities
nvar18	personalized_programs

Simple Statistics

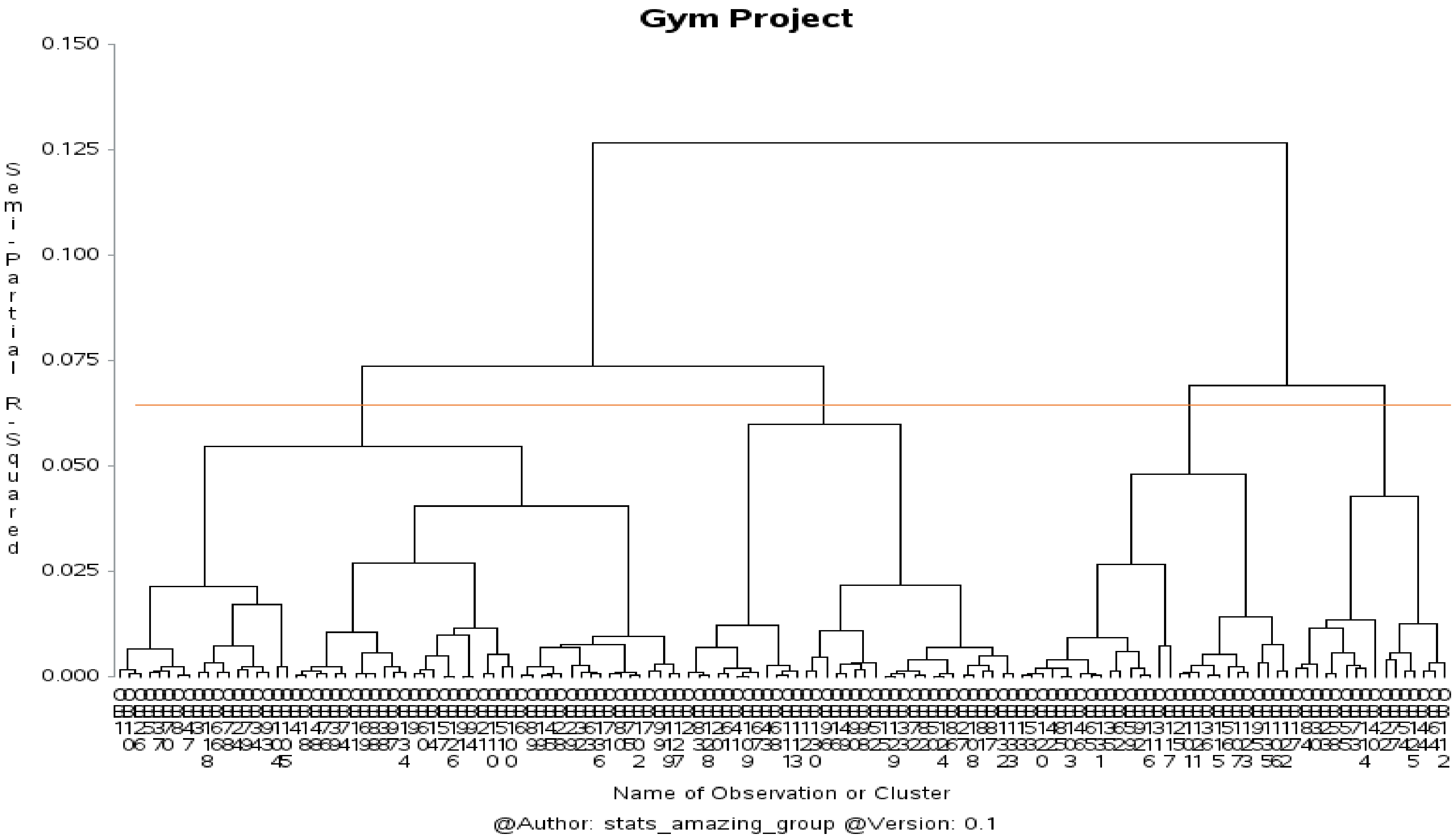
	nvar8	nvar9	nvar10	nvar11	nvar12	nvar13	nvar14	nvar15	nvar16	nvar17	nvar18
Mean	0.5816246620	-.1499848276	0.5449299199	-.0761469732	-.3026145570	-.1331843107	-.1357048369	0.6124883601	0.3403626311	-.5114024945	0.0469846234
StD	0.4992269406	0.6442917815	0.5626529510	0.7027896125	0.6388878328	0.7979720598	0.6123633071	0.4784562232	0.7527232391	0.6110602691	0.7105689498



Clustering method and dendrogram:

Using the variables treated for size effect and the first 6 principal components, the Ward method was used to construct 4 clusters that minimize the inter-cluster variance. In the figure below, the dendrogram and the cut-off point identifying the 4 clusters are visible. The accompanying table shows the frequency of observations in each cluster.

CLUSTER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	34	25.00	34	25.00
2	58	42.65	92	67.65
3	28	20.59	120	88.24
4	16	11.76	136	100.00



T-tests:

After segmenting the sample into four individual clusters, the mean of each attribute for each cluster was compared against a synthetic cluster made up by all the observations. This step was performed in order to identify the foundational characteristics of each cluster and is based on the assumption that those characteristics, the “partial” averages, should be distant from the “general” average of the fake cluster composed by all the data.

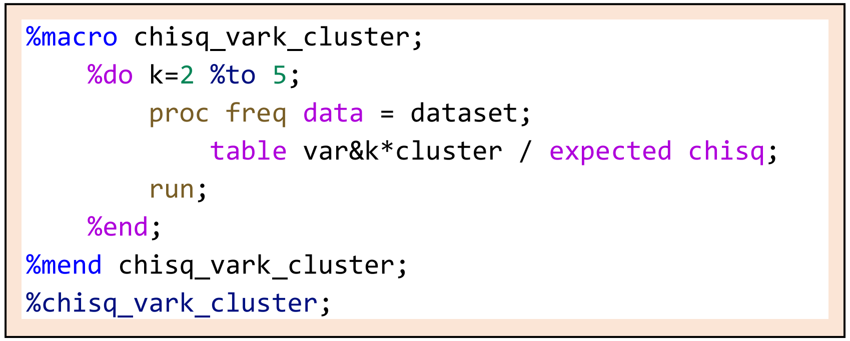
The next procedure was to perform a t-test over all of the scalable values for each of the clusters to determine if the individual active attributes were in fact significant to each of the clusters represented by the data. The t-value was used to determine whether a statistically significant difference exists between the mean of an individual cluster related to the entire dataset. Here the null hypothesis is given as no significant difference between the sample means. The results are illustrated in the table below.

descr	tvalue_clus 1	tvalue_clus 2	tvalue_clus 3	tvalue_clus 4	pvalue_clus 1	pvalue_clus 2	pvalue_clus 3	pvalue_clus 4
car_usage	3.22	-3.34	1.59	1.54	0.0019	0.0012	0.1183	0.1397
workout times	2.02	1.30	-2.57	-3.64	0.0488	0.1967	0.0138	0.0013
price	0.84	0.65	-1.83	0.59	0.4063	0.5201	0.0755	0.5596
installments	2.77	-1.97	0.91	-0.97	0.0076	0.0509	0.3677	0.3437
closeness home	1.34	2.63	-0.81	-4.27	0.1862	0.0095	0.4207	0.0005
closeness workplace	-4.40	1.32	2.55	-0.57	<.0001	0.1892	0.0142	0.5724
training_wfriends	-2.32	-1.02	1.59	2.72	0.0238	0.3103	0.1203	0.0139
courses	-2.51	-2.36	3.63	5.20	0.0150	0.0200	0.0007	<.0001
training_machines	-1.14	0.83	-2.15	2.07	0.2575	0.4065	0.0375	0.0537
time_flexibility	0.08	3.56	-4.21	2.35	0.9328	0.0005	0.0002	0.0259
park_availability	5.37	-4.41	3.35	1.60	<.0001	<.0001	0.0014	0.1252
amenities	-1.55	-0.89	-0.02	2.61	0.1250	0.3748	0.9846	0.0183
presonalized_program	-1.60	-0.12	0.60	2.90	0.1171	0.9020	0.5528	0.0082

- The orange rows relate to behaviour attributes
- Pink and green highlight significant t-values
- Blue represents borderline values

Chi-squared tests:

For the categorical variables such as gender, children, age and employment status a Chi squared test was implemented. The objective of this test is to check whether or not there is a significant deviation between the expected frequencies of the categorical variables with respect to the observed frequency. The code below was used in the SAS® analytic software to obtain the results.



```
%macro chisq_vark_cluster;
    %do k=2 %to 5;
        proc freq data = dataset;
            table var&k*cluster / expected chisq;
        run;
    %end;
%mend chisq_vark_cluster;
%chisq_vark_cluster;
```

Using the results obtained in the Chi squared tests, the frequencies and inferences were closely evaluated. A multivariate analysis of the categorical values for each individual cluster compared to a controlled group containing the other three clusters was also conducted. Shown below is an example of one of the contingency tables derived from the data. It represents the frequencies found in employment status for cluster 3 compared to the other clusters grouped together.

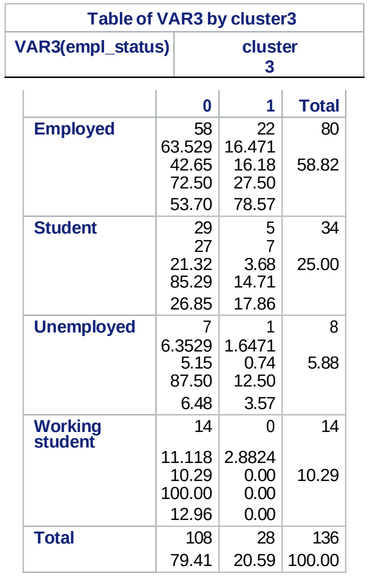


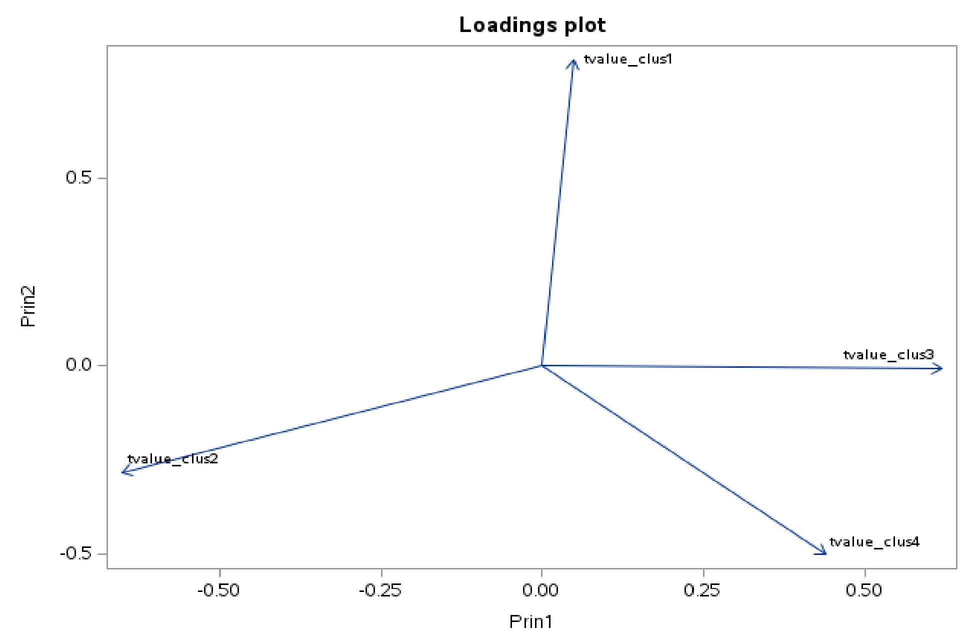
Table of VAR3 by cluster3

VAR3(empl_status)	cluster 3		
	0	1	Total
Employed	58 63.529 42.65 72.50 53.70	22 16.471 16.18 27.50 78.57	80 58.82
Student	29 27 21.32 85.29 26.85	5 7 3.68 14.71 17.86	34 25.00
Unemployed	7 6.3529 5.15 87.50 6.48	1 1.6471 0.74 12.50 3.57	8 5.88
Working student	14 11.118 10.29 100.00 12.96	0 2.8824 0.00 0.00 0.00	14 10.29
Total	108 79.41	28 20.59	136 100.00

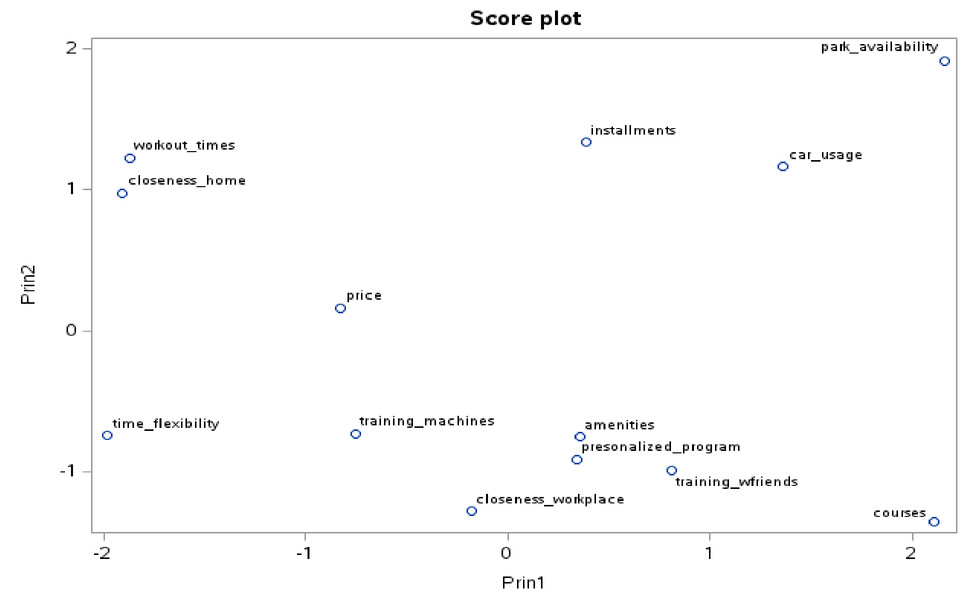
Component Pattern Plots:

The final process in the cluster analysis regards the creation of two plots from the values which emerge from the t-tests.
The first plot is a loadings plots and shows the impact of the variables over the first two principal components (i.e. the farther a variable is from the origin, the stronger the influence of the principal component it is on).
The second plot is the score plot and is created by multiplying the variables on which the principal components were calculated with the eigenvectors plotted in the loading graph; this multiplication casts the data in a new vectorial space where the new variables are linear combinations of the original eigenvectors.

If the score plot is compared to the loading plot, it demonstrates the impact of the eigenvectors over the projected data. Since the two plots were performed using the t-values, their interpretation is that the loadings plot shows the separation between the clusters' characteristics and the score plot shows how those cluster features are associated to the clusters.
For example, from the loading plot we know that the t-values for which cluster 1 is characterized point to the upper part of the graph. Those values correspond approximately to: car usage, the possibility to pay in installments and parking availability in the score plot. If we check for the t-values of those variables for cluster 1 in the t-test table, we will find the following values: 3.22, 2.77 and 5.37. So, by just looking at those two plots, an analyst can have a visual hint of the most important gym features for each cluster.



@Author: stats_amazing_group @Version: 0.1



@Author: stats_amazing_group @Version: 0.1

Cluster Description I & II:

CLUSTER I -

“ The Gym Buffs”

This cluster is composed of 34 respondents (25% of the total); it is composed of mainly young people between minor age and 35 years old (70%), it is the second cluster for presence of students (40%). This cluster is characterised by a strong presence of young people and students who have a job. People in this cluster are, in fact, very prone to pay in instalments. Among all the clusters they are the ones who seem to take working out the most seriously (t-value 2.02). They care less about being close to the workplace since they use a car (t-value 3.22), and therefore they seek available parking (it displays the highest t-value 5.37). These young people seemed to be more interested in weight lifting because they give less importance to courses and to working out with friends.

INTERPRETATION: These people workout for health and aesthetics and not to meet new people or friends, therefore the cluster is named “The Gym Buffs”. They are probably in the early stages of their working life so they probably don’t earn much money, and paying in instalments helps them train regularly without spending all at once.

CLUSTER II - “The Comfy Students”

The second cluster is the biggest one and it is composed of 58 people which account for 42.65% of the total respondents. This cluster is also composed of young people with a percentage of 83% minors to 35 year olds. 50% of people in the cluster are students and working students but, unlike cluster I, they seem to use the car less (t-value -3.34) therefore they do not give importance to the presence of car parks (it displays the lowest t-value -4.41). Since they don’t really use a car, closeness to home for them is very important (t-value 2.63) and they also seem to care a lot about time flexibility (t-value 3.56).

INTERPRETATION: These people seem to not have a car, therefore they care a lot for the gym to be close to their house, which is why this cluster is named “The Comfy Students”. Students and working students also have a tight schedule which translates into a bigger importance on time flexibility.

Cluster Description III & IV:

CLUSTER III - “The Workaholics”

The third cluster is composed of 28 people (accounting for the 20.59% of the total respondents). This cluster is prevalently composed of women (68%) and mostly employed. From this cluster on, an increase in the age of the components can be observed. The people in this cluster don’t really workout that much (the t-value observed is -2.57), they don’t deem training machines or time flexibility as very important (t-values are respectively -2.15 and -4.21). these people do care about closeness to workplace (he p-value registered is 2.55) and courses and parking (t-value 3.35)

INTERPRETATION: These people are at the age where they are either very busy with work or probably children (25% are parents) and their partner, thus they only workout if they have the time to, and they prefer going to courses and mostly close to their workplace, therefore they are named “The Workaholics”. They want to go to a gym that is more sustainable for their lifestyle.

CLUSTER IV -

“The Fancy Ones”

The fourth and last cluster is composed of 16 people, which are the remaining 11.76% of the respondents. The people in this cluster are mostly women (62%), are significantly older than the people in the first two clusters (it hosts the majority of the people from 46 years old+ out of all the respondents) and they are more likely to have children (31% of them are parents). Working out is not perceived as a priority for them, with an observed t-value of -3.64. These people seem to care more about the social aspects of the gym such as training with friends and courses. (t-values are 2.72 and 5.20). They are the only cluster that really cares about amenities, training machines and personalised programs (t-values of 2.61, 2.07 and 2.90)

INTERPRETATION: These people don’t really go to a gym for various reasons such as being busy with work or with children. This cluster has the highest percentage of unemployed people (50% out of all the respondents) which could be the mothers. If they had to go to a gym, they would prefer one with all the amenities and they would rather be helped by a professional since they might not know much about working out, therefore this cluster has been labelled “The Fancy Ones”.

Conclusion:

Physical fitness has become increasingly more important in the last decades and gyms provide an excellent option for many. What may have begun as a fad many years ago has now grown into a solid industry. However when considering what attracts potential members or clients to a gym, there are shown to be many factors. The study concludes that those who may be younger and looking for pure physical prowess are often not interested in extra services. Whereas some of the people who might be older or women, and less frequent exercisers, are often looking for a facility that offers all of the extras and has a more social atmosphere. Additionally work plays a large role in gym activity. The group with the highest number of employed people in effect do not frequently workout and when looking for a gym, proximity to workplace, availability of different courses and parking are important. This could be interesting information for an organization considering opening a gym in an industrial or office area. Another interesting division regards age and study. The majority of students and people under 35 seem to consider closeness to home and flexibility in operating hours to be more of importance. Perhaps this also reflects a growing trend for many businesses to expand operating hours in order to accommodate an ever-evolving society, requiring more flexible schedules. Finally, what people are looking for at the gym is greatly determined by a set of commonalities that, in some sense, binds these people together. These commonalities that create groups, or clusters, that lead to a greater understanding of target markets and how to satisfy them. This is valuable information to have in a very competitive business world.