IMSE 514 – MULTIVARIATE STATISTICS
HOMEWORK 5

SURESH OOTY

Background:

The actual data collected hints that it was a survey done among students that learn/deal with Mathematics, Statistics & use Computers and SPSS. The collected responses for a set of 23 questions were one of 5 equal choices. The categorical choices are

- Strongly agree
- Agree
- Neither
- Disagree
- Strong disagree

A high level observation on the set of the questions & the choices given hint that the data could lead a set of interesting clusters.

Observation of data:

		Statistics		Computer		Math		SPSS	
Subject		Don't like	Like	don't	Like	don't	like	don't	like
Statistics makes me cry	V1	x							
My friends will think I'm stupid for not being able to cope with SPSS	V2							x	
Standard deviations excite me	V3		x						
I dream that Pearson is attacking me with correlation coefficients	V4	x							
I don't understand statistics	V5	x							
I have little experience of computers	V6			x					
All computers hate me	V7			x					
I have never been good at mathematics	V8					x			
My friends are better at statistics than me	V9	x							
Computers are useful only for playing games	V10			x					
I did badly at mathematics at school	V11					x			
People try to tell you that SPSS makes statistics easier to understand but it doesn't	V12							x	
I worry that I will cause irreparable damage because of my incompetence with computers	V13			x					
Computers have minds of their own and deliberately go wrong whenever I use them	V14			x					
Computers are out to get me	V15			x					
I weep openly at the mention of central tendency	V16	x							
I slip into a coma whenever I see an equation	V17					x			
SPSS always crashes when I try to use it	V18							x	
Everybody looks at me when I use SPSS	V19								x
I can't sleep for thoughts of Eigen vectors	V20					x			
I wake up under my duvet thinking that I am trapped under a normal distribution	V21	x							
My friends are better at SPSS than I am	V22							x	
If I'm good at statistics my friends will think I'm a nerd	V23	x							

Data Preparation:

The data could be converted as equivalent factors as in the table below.

• Strongly agree	5
• Agree	4
• Neither	3
• Disagree	2
• Strong disagree	1

When the correlation (Pearson's) was constructed using

```
Co_old<-corr(as.matrix(old), type="pearson")
Cr_old<-cbind.data.frame(Co_old$r)
```

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19	v20	v21	v22	v23
v1	1	-0.09872	-0.33665	0.43586	0.40244	0.216734	0.305365	0.330738	-0.09234	0.213682	0.356786	0.345381	0.354646	0.33788	0.245753	0.498618	0.370551	0.347118	-0.18901	0.213898	0.329153	-0.10441	-0.00448
v2	-0.09872	1	0.31839	-0.11186	-0.11935	-0.07421	-0.15917	-0.04962	0.314641	-0.084	-0.14383	-0.19487	-0.14274	-0.1647	-0.165	-0.16755	-0.087	-0.16389	0.203297	-0.20159	-0.20462	0.230875	0.099678
v3	-0.33665	0.31839	1	-0.38046	-0.31031	-0.22674	-0.38195	-0.25863	0.299804	-0.19339	-0.35064	-0.40995	-0.31792	-0.37076	-0.3124	-0.41865	-0.32737	-0.37523	0.341574	-0.32483	-0.41719	0.203657	0.150207
v4	0.43586	-0.11186	-0.38046	1	0.400672	0.278202	0.408615	0.349429	-0.12455	0.21581	0.368657	0.441647	0.344292	0.35081	0.334231	0.415867	0.382739	0.382001	-0.18598	0.242918	0.410293	-0.09838	-0.03382
v5	0.40244	-0.11935	-0.31031	0.400672	1	0.25746	0.339392	0.268627	-0.0957	0.258209	0.297829	0.346743	0.301822	0.315338	0.261372	0.394918	0.310417	0.322091	-0.16532	0.199669	0.334615	-0.13254	-0.04166
v6	0.216734	-0.07421	-0.22674	0.278202	0.25746	1	0.513581	0.222832	-0.11264	0.32223	0.328071	0.312509	0.466405	0.402244	0.359893	0.244339	0.282261	0.513322	-0.16675	0.100925	0.272333	-0.16514	-0.06869
v7	0.305365	-0.15917	-0.38195	0.408615	0.339392	0.513581	1	0.297497	-0.1283	0.283723	0.344748	0.422986	0.442119	0.440703	0.391367	0.388545	0.390743	0.500867	-0.26912	0.220954	0.483004	-0.1682	-0.07029
v8	0.330738	-0.04962	-0.25863	0.349429	0.268627	0.222832	0.297497	1	0.015733	0.158609	0.629298	0.251986	0.314247	0.28059	0.299686	0.321494	0.59014	0.279744	-0.15948	0.175151	0.295718	-0.07917	-0.05024
v9	-0.09234	0.314641	0.299804	-0.12455	-0.0957	-0.11264	-0.1283	0.015733	1	-0.13419	-0.11552	-0.16739	-0.16744	-0.1215	-0.18657	-0.18887	-0.03682	-0.14958	0.249312	-0.15865	-0.13594	0.256846	0.170774
v10	0.213682	-0.084	-0.19339	0.21581	0.258209	0.32223	0.283723	0.158609	-0.13419	1	0.271437	0.245826	0.301967	0.254687	0.295234	0.290586	0.218322	0.292503	-0.12723	0.084065	0.193136	-0.13091	-0.06192
v11	0.356786	-0.14383	-0.35064	0.368657	0.297829	0.328071	0.344748	0.629298	-0.11552	0.271437	1	0.335295	0.423165	0.32532	0.364827	0.369078	0.586835	0.373414	-0.19965	0.255337	0.346434	-0.16199	-0.08637
v12	0.345381	-0.19487	-0.40995	0.441647	0.346743	0.312509	0.422986	0.251986	-0.16739	0.245826	0.335295	1	0.488713	0.432704	0.331799	0.408059	0.332694	0.492965	-0.26666	0.298026	0.440638	-0.16729	-0.04643
v13	0.354646	-0.14274	-0.31792	0.344292	0.301822	0.466405	0.442119	0.314247	-0.16744	0.301967	0.423165	0.488713	1	0.449786	0.342197	0.358378	0.408377	0.532937	-0.22697	0.203963	0.374431	-0.19536	-0.05298
v14	0.33788	-0.1647	-0.37076	0.35081	0.315338	0.402244	0.440703	0.28059	-0.1215	0.254687	0.32532	0.432704	0.449786	1	0.380115	0.418418	0.353742	0.498306	-0.25406	0.225922	0.399389	-0.16984	-0.04847
v15	0.245753	-0.165	-0.3124	0.334231	0.261372	0.359893	0.391367	0.299686	-0.18657	0.295234	0.364827	0.331799	0.342197	0.380115	1	0.454279	0.373102	0.34287	-0.2098	0.206256	0.299716	-0.16791	-0.06201
v16	0.498618	-0.16755	-0.41865	0.415867	0.394918	0.244339	0.388545	0.321494	-0.18887	0.290586	0.369078	0.408059	0.358378	0.418418	0.454279	1	0.409763	0.421979	-0.26705	0.26514	0.420543	-0.15579	-0.08152
v17	0.370551	-0.087	-0.32737	0.382739	0.310417	0.282261	0.390743	0.59014	-0.03682	0.218322	0.586835	0.332694	0.408377	0.353742	0.373102	0.409763	1	0.375607	-0.16288	0.20523	0.363491	-0.12629	-0.09167
v18	0.347118	-0.16389	-0.37523	0.382001	0.322091	0.513322	0.500867	0.279744	-0.14958	0.292503	0.373414	0.492965	0.532937	0.498306	0.34287	0.421979	0.375607	1	-0.25663	0.23518	0.430104	-0.15983	-0.08042
v19	-0.18901	0.203297	0.341574	-0.18598	-0.16532	-0.16675	-0.26912	-0.15948	0.249312	-0.12723	-0.19965	-0.26666	-0.22697	-0.25406	-0.2098	-0.26705	-0.16288	-0.25663	1	-0.24859	-0.2749	0.233923	0.122434
v20	0.213898	-0.20159	-0.32483	0.242918	0.199669	0.100925	0.220954	0.175151	-0.15865	0.084065	0.255337	0.298026	0.203963	0.225922	0.206256	0.26514	0.20523	0.23518	-0.24859	1	0.467704	-0.0997	-0.03467
v21	0.329153	-0.20462	-0.41719	0.410293	0.334615	0.272333	0.483004	0.295718	-0.13594	0.193136	0.346434	0.440638	0.374431	0.399389	0.299716	0.420543	0.363491	0.430104	-0.2749	0.467704	1	-0.12902	-0.06766
v22	-0.10441	0.230875	0.203657	-0.09838	-0.13254	-0.16514	-0.1682	-0.07917	0.256846	-0.13091	-0.16199	-0.16729	-0.19536	-0.16984	-0.16791	-0.15579	-0.12629	-0.15983	0.233923	-0.0997	-0.12902	1	0.230369
v23	-0.00448	0.099678	0.150207	-0.03382	-0.04166	-0.06869	-0.07029	-0.05024	0.170774	-0.06192	-0.08637	-0.04643	-0.05298	-0.04847	-0.06201	-0.08152	-0.09167	-0.08042	0.122434	-0.03467	-0.06766	0.230369	1

Some correlations could be noted among the variables that are about ~0.62 between v8 & v11. And the second one ~ 0.58 between v8 & v17 and v17 & v11.

I have never been good at mathematics - V8
I did badly at mathematics at school - V11
I slip into a coma whenever I see an equation - V17

All of them deal with Math.

However, to calculate the dissimilarities between the categorical values need to be reduced from 5 to binary values. Note: When the steps were continued further by calculating the dissimilarity matrix and constructing the hierarchical clusters, the dendrogram became crowded.

A decision was made (after some research in internet [1]) to reduce the dimensions of the categorical values to 0, 1 and neither. This is done as follows.

• Strongly agree	Y
• Agree	
• Neither	neither
• Disagree	N
• Strong disagree	

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12
1	y	y	n	y	y	y	neither	y	y	y	y	y
2	y	y	n	neither	y	y	y	y	n	y	y	neither
3	y	neither	y	y	n	y	y	y	y	y	neither	neither
4	neither	y	y	n	neither	neither	n	y	y	n	y	y
5	y	y	neither	y	y	neither	neither	y	n	y	y	neither
6	y	y	neither	y	n	n	n	y	n	neither	y	n
7	y	neither	neither	y	y	y	y	y	neither	y	y	y
8	y	y	neither	y	y	y	y	y	n	y	y	neither
9	neither	neither	y	n	n	neither	n	n	neither	neither	n	n
10	y	n	n	neither	y	y	y	y	neither	y	y	neither
11	y	y	n	y	y	y	y	y	n	y	y	neither
12	y	y	neither	neither	n	neither	neither	y	neither	y	y	neither
13	neither	y	neither	n	neither	y	neither	neither	y	neither	neither	n
14	y	y	y	y	y	y	neither	y	y	neither	y	n
15	y	y	neither	n	y	y	neither	y	y	neither	y	neither
16	neither	y	y	y	y	y	y	y	y	neither	y	neither

<A snapshot of modified data>

Execution:

Once the dataset was manually modified as stated above. The following set of R codes were used to achieve this.

```
library(ade4)
disj<-acm.disjonctif(new)
```

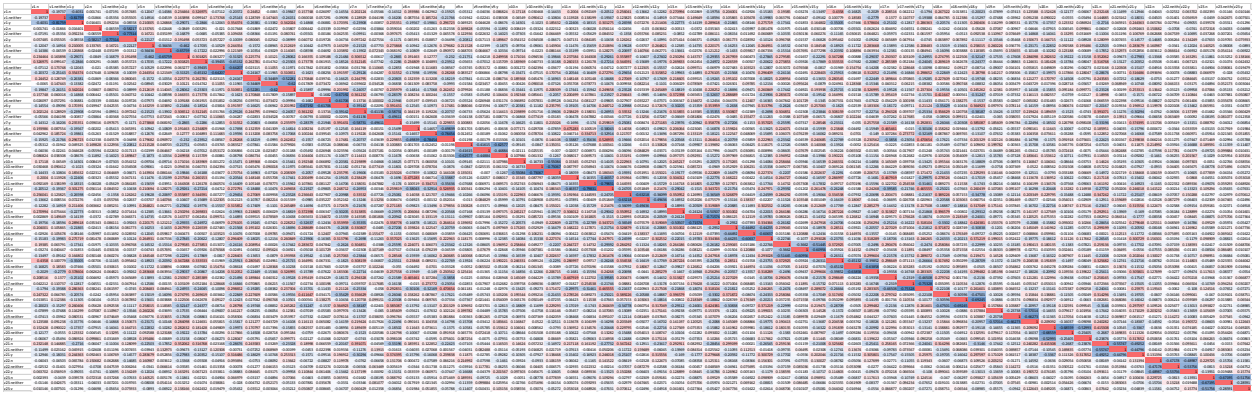
cityblock_for_y_n_neither.R | new | disj | cr | cr_old

Filter

	v1.n	v1.neither	v1.y	v2.n	v2.neither	v2.y	v3.n	v3.neither	v3.y	v4.n	v4.neither	v4.y	v5.n
1	0	0	1	0	0	1	1	0	0	0	0	1	0
2	0	0	1	0	0	1	1	0	0	0	1	0	0
3	0	0	1	0	1	0	0	0	1	0	0	1	1
4	0	1	0	0	0	1	0	0	1	1	0	0	0
5	0	0	1	0	0	1	0	1	0	0	0	1	0
6	0	0	1	0	0	1	0	1	0	0	0	1	1
7	0	0	1	0	1	0	0	1	0	0	0	1	0
8	0	0	1	0	0	1	0	1	0	0	0	1	0
9	0	1	0	0	1	0	0	0	1	1	0	0	1
10	0	0	1	1	0	0	1	0	0	0	1	0	0
11	0	0	1	0	0	1	1	0	0	0	0	1	0
12	0	0	1	0	0	1	0	1	0	0	1	0	1
13	0	1	0	0	0	1	0	1	0	1	0	0	0
14	0	0	1	0	0	1	0	0	1	0	0	1	0

This separates the y, n & neither in to separate columns (69 of them), with 1s representing the existence of column value (y or n or neither) for that subject.

A correlation test reveals that hardly any relationship, which is true as the data has been transposed from values to columns. (Red cells = 1; Blue cells = -ve)



The reference paper [1] uses Dice coefficient

Dice coefficient. Squared difference between the dummy coding 0/1 for each category of variables. Square of the Euclidean distance.

$$\delta^2_{jj'} = \frac{1}{2}\sum_{i=1}^{n}\left(m_{ij} - m_{ij'}\right)^2$$

i is the individual n°i
j is the jth category
m_{ij} is an indicator for the jth category

Which made the dissimilarity more visible for a categorical values.
Constructing Dice Index matrix in R code:

cityblock_for_y_n_neither.R | new | disj | cr | cr_old

Source on Save

```
1  #Dice's index
2  dice <- function(m1,m2){
3    return(0.5*sum((m1-m2)^2))
4  }
5  #Dice's index matrix
6
7  d2 <- matrix(0,ncol(disj),ncol(disj))
8  for (j in 1:ncol(disj)){
9    for (jprim in 1:ncol(disj)){
10     d2[j,jprim] <- dice(disj[,j],disj[,jprim])
11   }
12 }
13 colnames(d2) <- colnames(disj)
14 rownames(d2) <- colnames(disj)
15
16 #transform the matrix in a R 'dist' class
17 d <- as.dist(sqrt(d2))
18
```

D contains the dissimilarity structure the could be used in hclust [2]

When the hclust of tried on the given dataset of 69 columns, A slightly overcrowded hierarchical dendrogram structure showed up.

```
x<- hclust(d, method =”ward.D2”)
Plot(x)
```

The overcrowded dendrogram did not show meaningful structure with bunch of “neither” leaf nodes sticking together. To make the structure lean and meaningful, the “neither” columns were dropped out of the dataset and the study was done from beginning.

```
disjnew<-disj[cbind(1,3,4,6,7,9,10,12,13,15,16,18,19,21,22,24,25,27,28,30,31,33,34,36,37,39,40,42,43,45,46,48,49,51,52,54,55,57,58,60,61,63,64,66,67,69…
disjnew<-cbind.data.frame(disj[c(1,3,4,6,7,9,10,12,13,15,16,18,19,21,22,24,25,27,28,30,31,33,34,36,37,39,40,42,43,45,46,48,49,51,52,54,55,57,58,60,61,6…
```

```
20  # after removing neither category
21  d3 <- matrix(0,ncol(disjnew),ncol(disjnew))
22  for (j in 1:ncol(disjnew)){
23    for (jprim in 1:ncol(disjnew)){
24      d3[j,jprim] <- dice(disjnew[,j],disjnew[,jprim])
25    }
26  }
27  colnames(d3) <- colnames(disjnew)
28  rownames(d3) <- colnames(disjnew)
29  #transform the matrix in a R ‘dist’ class
30  d3<- as.dist(sqrt(d3))
31
```

With this reduced Dice’s index matrix, the hierarchical cluster was tried with methods

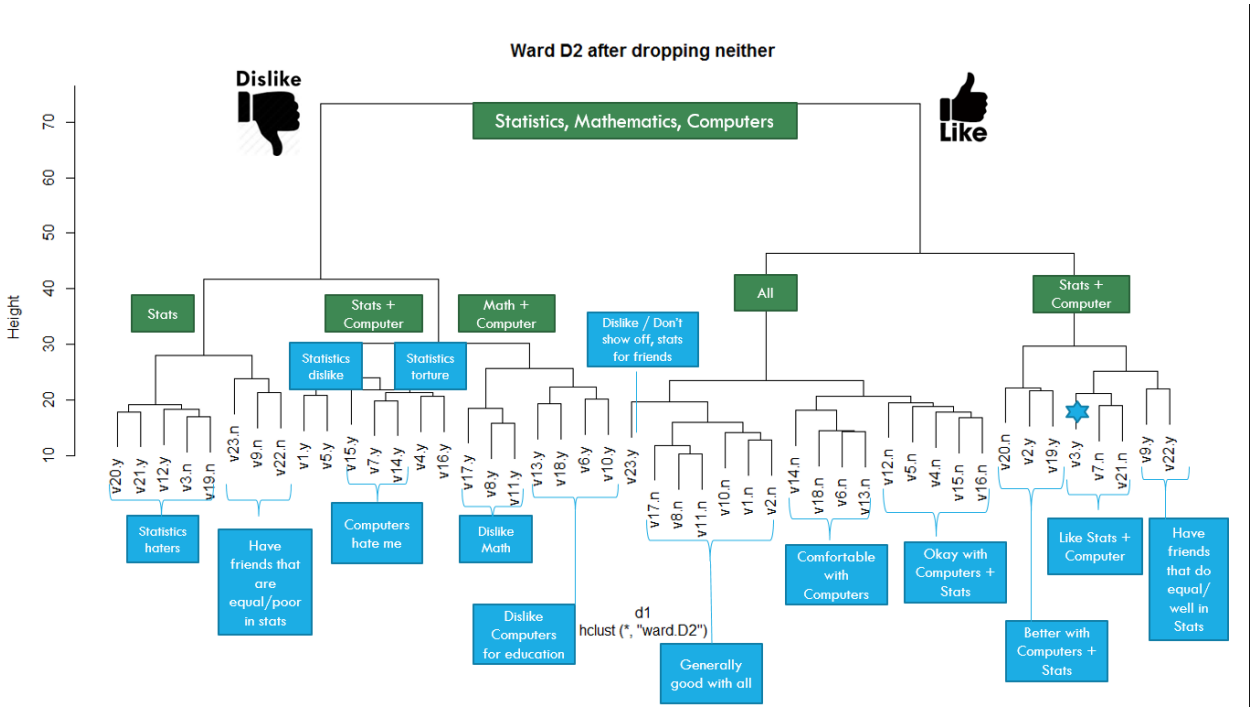
- method = “ward.D2”
- method = “complete”
- method = “average”

“Single” method was observed as to be too crowded given the high number of questions.

Note that the variables V1 to V23 are as per the table shown in the “observation of dataset”
V1.y = the response for “Statistics makes me cry” is either Agree or Strongly Agree
V1.n = the response for “Statistics makes me cry” is either Disagree or Strongly disagree

Visualization:

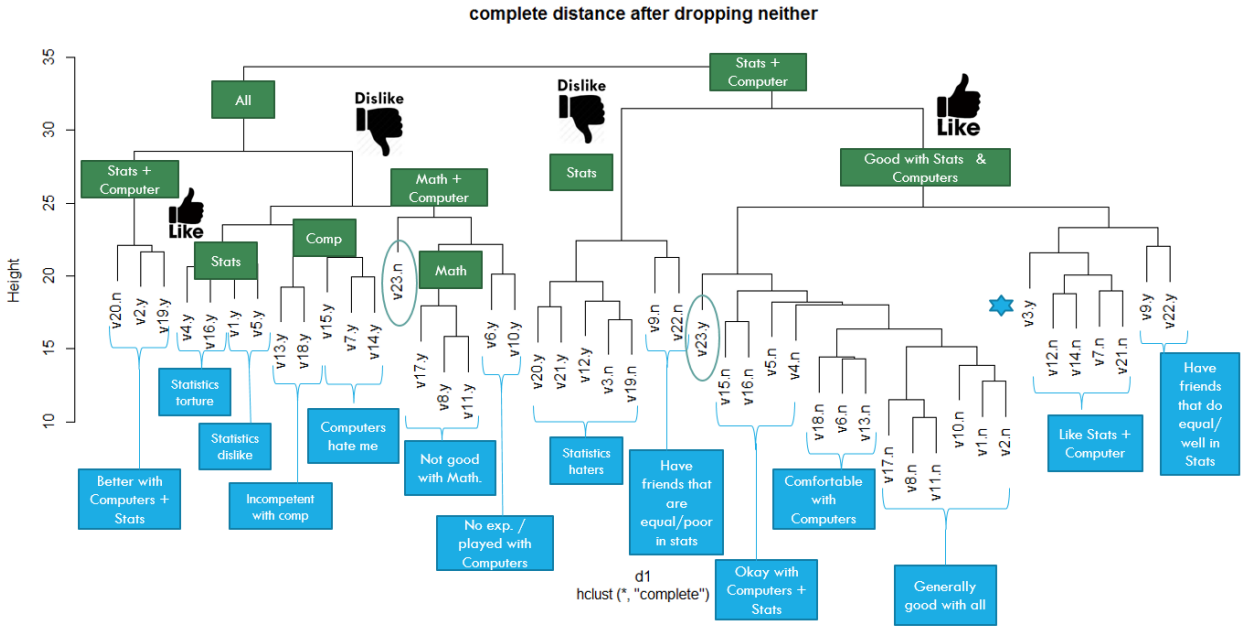
Hierarchical cluster using Ward D2 method:



Observations:

- The cluster divides largely such that factors divide the students in to “Like” and “Dislike” categories. Using which, the students could be categorized grouped further sub-groups.
- From left to right, the distribution is “Hate” to “Dislike” to “Okay” to “Like”
- Note that V3.y (*Standard deviations excite me*) and V19.y(*Everybody looks at me when I use SPSS*) are the placed together which are the factors that could conceived as few questions termed in positive language (someone who likes studies very much)
- In this method, at the level 2 (where green boxes end) makes a logical means to group the students based on their INTEREST towards the considered disciplines.

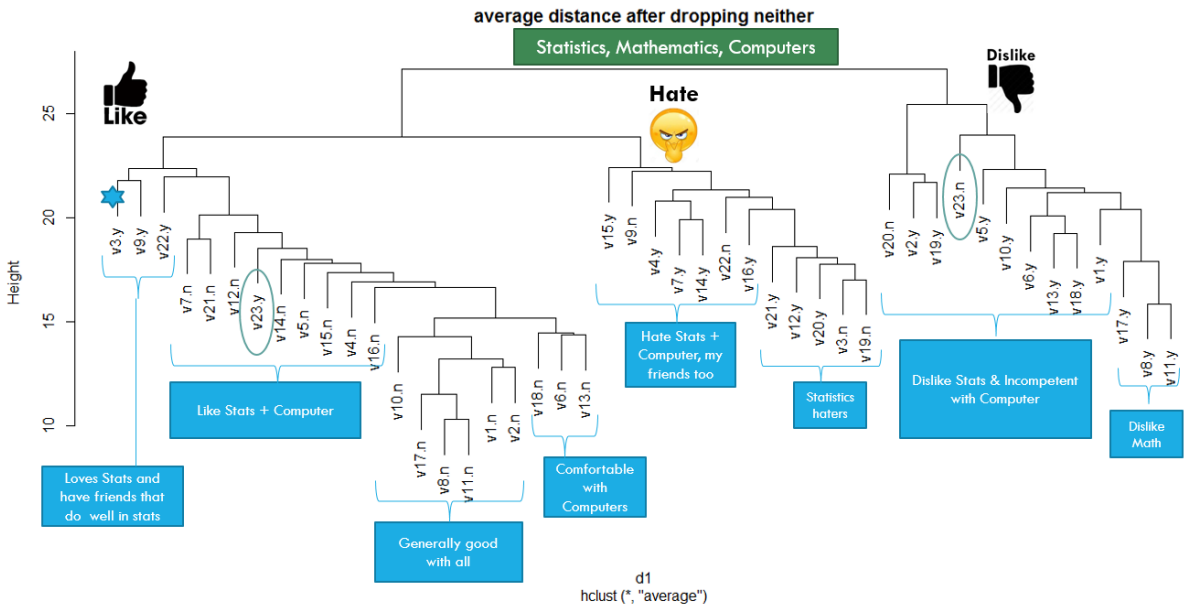
Hierarchical cluster using complete method:



Observations:

- The first impression is that the “Like” & “Dislike” groups are arranged in the middle between two branches. This does not have the “left to right” arrangement noted in “ward.D2” method.
- The broken down tree separates the first level tree in to “All” subjects against “Stats + Computer” and then further breaks each group in to “Like” & “Dislike” groups.
- It is noted that the two “Dislike” branches under different groups, are arranged closer to each other.
- The complete distance groups the clusters first and then breaks in to sub-groups based on the farthest distances.
- It is also noted that V23.y/n (*If I'm good at statistics my friends will think I'm a nerd*) is somewhat misplace in the group. A disadvantage when large number of factors are analyzed using this approach.
- Even if the tree is cut at level 3, it becomes difficult to make sense out of this structure. But a significance of study on Statistics & Computers could be noted from the tree structure, based on the number of questions around these two subjects and they have less dissimilar outcomes.

Hierarchical cluster using average method:



Observations:

- The average distance breaks the tree structure in to three groups. This was not clearly observed on the above two approaches. A new Interest level is identified (I call them haters!)
- There is no left to right order that was observed in ward.D2 method. An average distance between clusters would have danced around like a bubble to grow in to 3 major branches.
- It could be see that under “Like” branch and “Hate” branch, there is a set of questions that could be used to group a set of subjects (students) that hang out with likeminded friends. This was not so apparently visible in the other two approaches.
- If the other two methods brought the “disciplines” to lime light, average method brought the “behavior” or interest level of the subjects.

Conclusion:

The subjects could be grouped based on the average method to identify the “Likers”, “Disliker” (might need some coaching) and “Haters” (might need counselling).

Also note that the variables that were found to be correlated V8, V11 & V17 (all ‘n’s) have come together in all the three tree structures. ☺

References

[1] http://eric.univ-lyon2.fr/~ricco/cours/slides/en/classif_variables_quali.pdf
[2] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html
[3] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/dist.html