ASSIGNMENT SUBMISSION FORM

This will be the first page of your assignment

Course Name : **Unsupervised Machine Learning - 1**
Assignment Title : **Individual Assignment 2**
Submitted by : **Unnati Khinvasara**

Student Name	PG ID
Unnati Khinvasara	12120097

ISB Honour Code

- I will represent myself in a truthful manner.
- I will not fabricate or plagiarise any information with regard to the curriculum.
- I will not seek, receive or obtain an unfair advantage over other students.
- I will not be a party to any violation of the ISB Honour Code.
- I will personally uphold and abide, in theory and practice, the values, purpose and rules of the ISB Honour Code.
- I will report all violations of the ISB Honour Code by members of the ISB community.
- I will respect the rights and property of all in the ISB community.
- I will abide by all the rules and regulations that are prescribed by ISB.

Note: Lack of awareness of the ISB Honour Code is never an excuse for a violation. Please go through the Honour Code in the student handbook, understand it completely. Please also pay attention to the following points:

- Please do not share your assignment with your fellow students under any circumstances if the Honour Code scheme prohibits it. The HCC considers both parties to be guilty of an Honour Code violation in such circumstances.
- If the assignment allows you to refer to external sources, please make sure that you cite all your sources. Any material that is taken verbatim from an external source (website, news article etc.) must be in quotations. A much better practice is to paraphrase the source material (it still must be cited).

(Please start writing your assignment below)

Data Background

We have been given dataset of 178 Wines along with its 14 variables. The first variable is the Wine Cultivar Group which categorises the Wines into 3 Groups and the rest 13 variables are various chemical components of the wine. The list of the other 13 variables in order is as follows –

1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10) Color intensity
11) Hue
12) OD280/OD315 of diluted wines
13) Proline

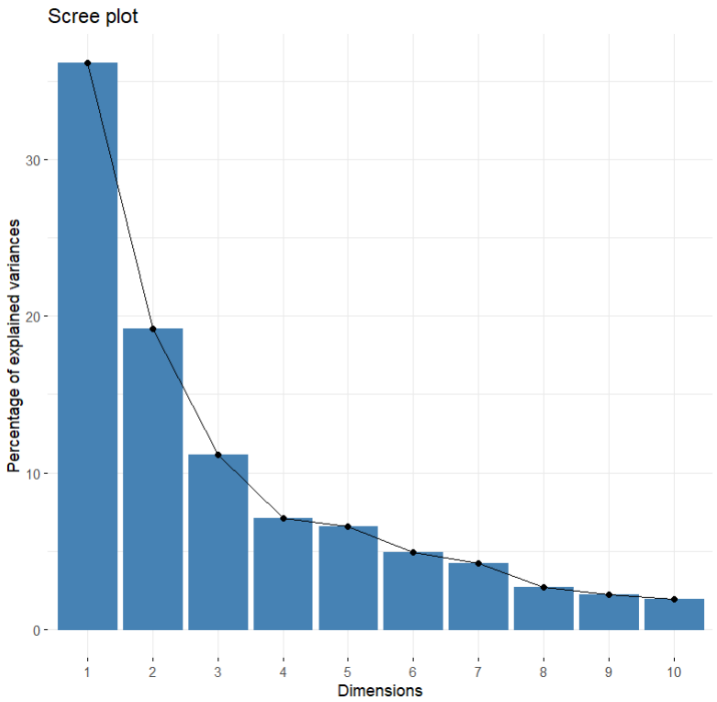
Q(a) - Enumerate the insights you gathered during your PCA exercise.

We start off by exploring the dataset and realized that the Wine Cultivars variable is categorical and hence not required in the PCA exercise. After removing the same, we conducted the PCA Exercise.

We start with checking the variances captured by the PCA –

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6     Comp.7
Standard deviation     2.1692972 1.5801816 1.2025273 0.9586313 0.92370351 0.80103498 0.74231281
Proportion of Variance 0.3619885 0.1920749 0.1112363 0.0706903 0.06563294 0.04935823 0.04238679
Cumulative Proportion  0.3619885 0.5540634 0.6652997 0.7359900 0.80162293 0.85098116 0.89336795
                           Comp.8     Comp.9    Comp.10    Comp.11    Comp.12     Comp.13
Standard deviation     0.59033665 0.53747553 0.50090167 0.47517222 0.41081655 0.321524394
Proportion of Variance 0.02680749 0.02222153 0.01930019 0.01736836 0.01298233 0.007952149
Cumulative Proportion  0.92017544 0.94239698 0.96169717 0.97906553 0.99204785 1.000000000
> |
```

We can graphically represent the same as follows –

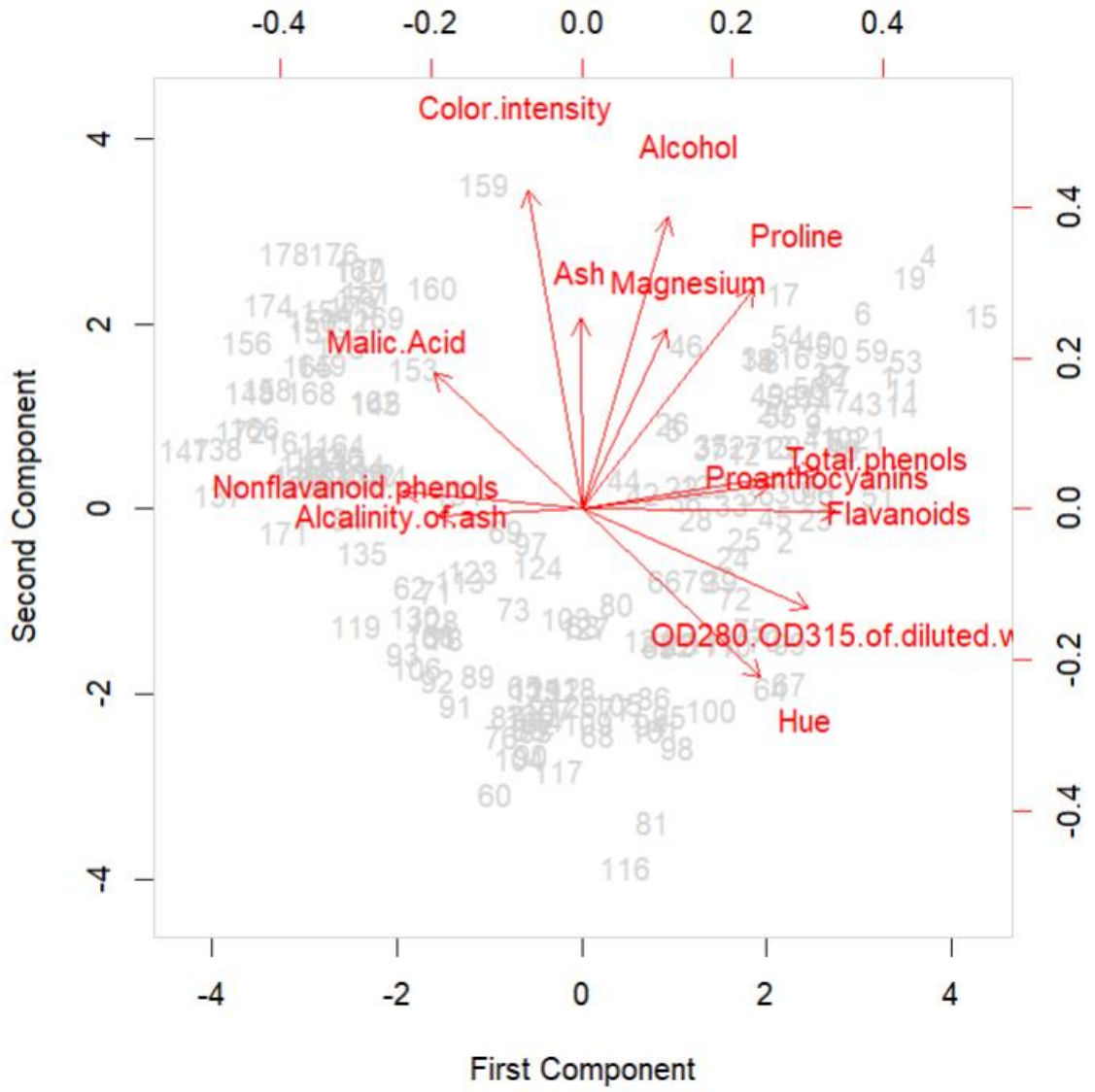


From this, we can infer that PC1 captures 36% of the data, PC2 captures 19% of the data, PC3 covers 11% of the data and so on.

Further, we can also note that to cover ~90% of the data, PC1 to PC7 should be used together.

Further, it should also be noted that PC1 covers 2.16 standard deviations, PC2 covers 1.58 SD and it keeps dropping from thereon.

Also, since we can see that PC1 & PC2 covers almost 55% of the data, we have tried to plot these down to check the direction of points and the vectors.



The relative location of the points can be interpreted. Further, both the direction and length of the Eigenvectors can be interpreted.

Observations on the right side of the graph have a similar value for Total Phenol / Proanthocyanin / Flavonoid content. Similarly, observation on the left have similar values for Nonflav phenols and alkalinity of ash.

Further, with respect to the variables, we can understand the direction to infer that Total Phenol / Proanthocyanin / Flavanoids are positively impacted by increase in PC1 whereas the other side of nonflav phenols and alkalinity of ash are negatively impacted. Both are important contributors for PC1. Similarly variables in the direction of upwards and downward are important contributors for PC2.

It should be noted that there is high correlation between variables that are grouped together.

Further, one point to note that is Ash is almost uncorrelated with other variables.

Further, the above understanding can be cross checked with the loadings (i.e. weights) given below. Higher the weight, higher the impact of the variable on the PC.

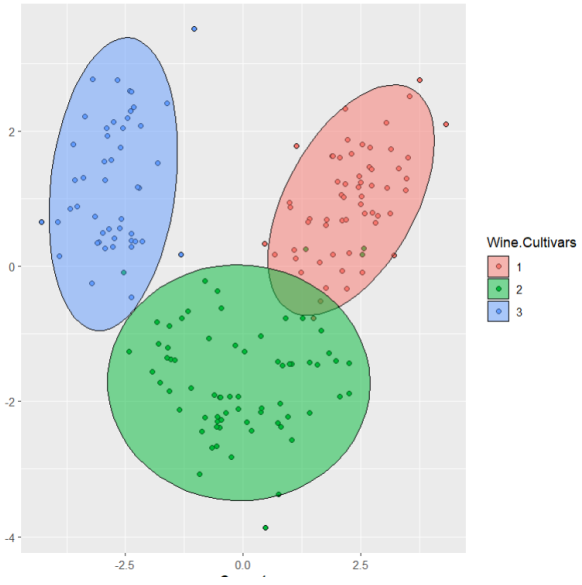
```
Loadings:
                               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
Alcohol                         0.144  0.484  0.207         0.266  0.214         0.396  0.509  0.212
Malic.Acid                     -0.245  0.225        -0.537         0.537 -0.421                -0.309
Ash                                    0.316 -0.626  0.214  0.143  0.154  0.149 -0.170 -0.308
Alcalinity.of.ash              -0.239        -0.612                -0.101  0.287  0.428  0.200
Magnesium                       0.142  0.300 -0.131  0.352 -0.727        -0.323 -0.156  0.271
Total.phenols                   0.395        -0.146 -0.198  0.149                -0.406  0.286 -0.320
Flavanoids                      0.423        -0.151 -0.152  0.109                -0.187        -0.163
Nonflavanoid.phenols           -0.299        -0.170  0.203  0.501 -0.259 -0.595 -0.233  0.196  0.216
Proanthocyanins                 0.313        -0.149 -0.399 -0.137 -0.534 -0.372  0.368 -0.209  0.134
Color.intensity                       0.530  0.137                -0.419  0.228                -0.291
Hue                             0.297 -0.279         0.428  0.174  0.106 -0.232  0.437         -0.522
OD280.OD315.of.diluted.wines    0.376 -0.164 -0.166 -0.184  0.101  0.266                0.137  0.524
Proline                         0.287  0.365  0.127  0.232  0.158  0.120         0.120 -0.576  0.162
                               Comp.11 Comp.12 Comp.13
Alcohol                         0.226   0.266
Malic.Acid                             -0.122
Ash                             0.499          -0.141
Alcalinity.of.ash              -0.479
Magnesium
Total.phenols                  -0.304   0.304  -0.464
Flavanoids                              0.832
Nonflavanoid.phenols           -0.117           0.114
Proanthocyanins                 0.237          -0.117
Color.intensity                        -0.604
Hue                                    -0.259
OD280.OD315.of.diluted.wines           -0.601  -0.157
Proline                        -0.539
```

Q(b) - What are the social and/or business values of those insights, and how the value of those insights can be harnessed—enumerate actionable recommendations for the identified stakeholder in this analysis?

The value we can derive from the above insight is in terms of the chemical composition of the wines and which chemical seem to be correlated to each other. This can help the business in understanding that say colour / hue of the wine has no correlation with other chemicals such as total phenols, flavonoids, etc. It can also aid in understanding that chemicals like flavonoids / total phenols / OD280, etc are major impactors for the composition of the wine. And hence wines are sensitive to slight changes in these chemicals.

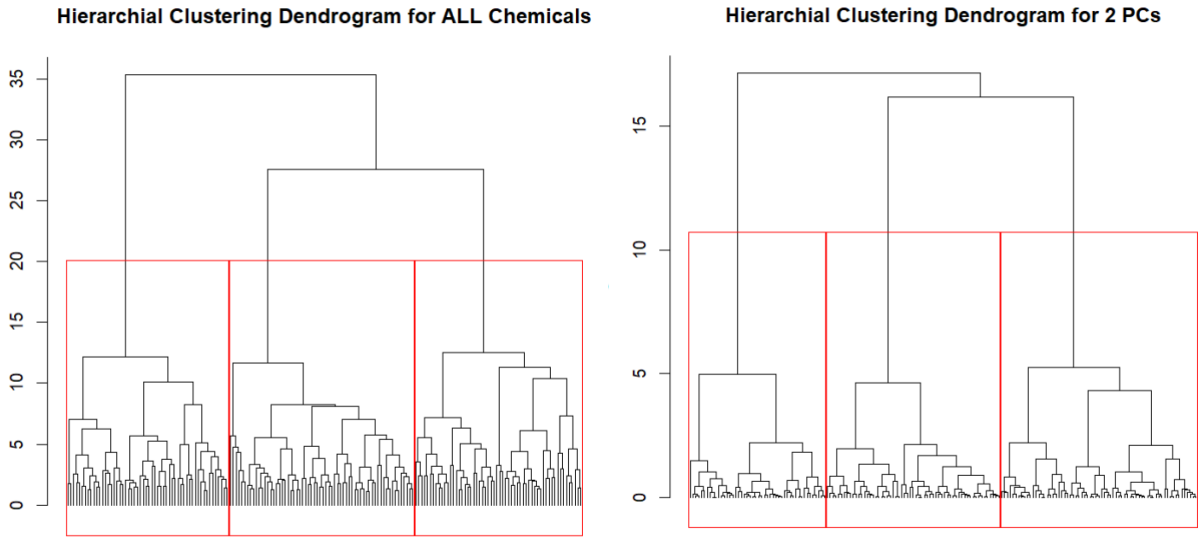
Further, it can also help the business in categorizing the wines based on the chemical composition and content. If any chemical is not very impactful (say like Ash), the business can take a call on the quality of ash to be sourced and view at cost cutting of such chemicals.

Further, for the consumer, this analysis should be able to help understand the chemical component which majorly impact different type of wine and they can then further look into the health risks or nutrition impact of the wine type and decide which one to consume accordingly.



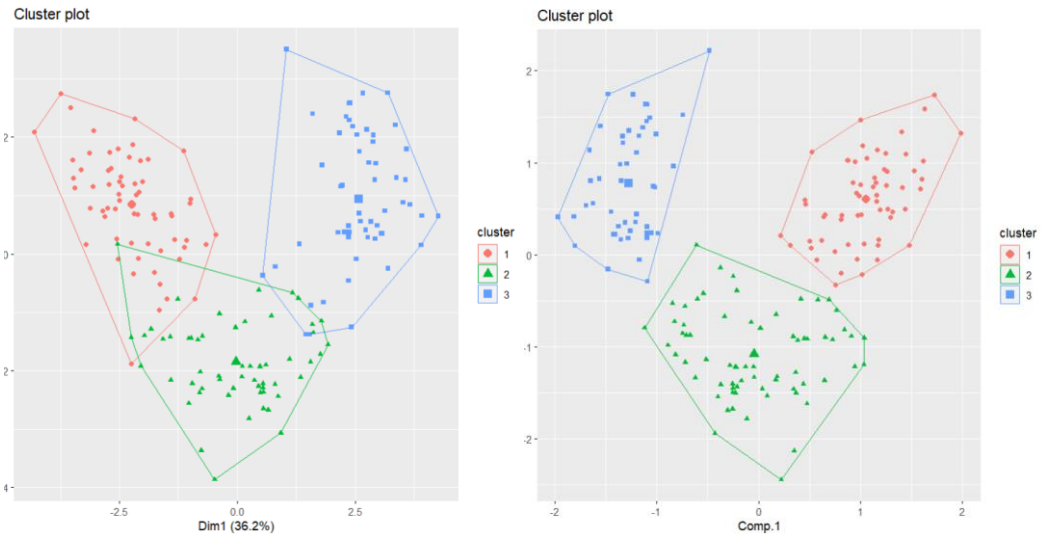
Q(c) - Any more insights you come across during the clustering exercise?

The following are the dendrograms from the clustering exercise –



We can see that with using all 13 chemicals and even with only 2 PCs (which cover 55% of the total var), the clusters which clearly form are 3. If we drill down, then probably looking at the heights of the dendrogram, there would be differences. But on the overall, we have 3 major clusters.

If we try visually plotting, we see a slight overlap in the 13 chemical one as against 2 PC clustering. This is because 2 PC clustering only takes into account 55% of the data whereas the other one uses all chemicals.



(Plotting for all 13 chemical clusters) *(Plotting for 2 PCs clusters)*

From this we can infer that there are minor changes in the clusters formed via both.

Q(d) - Are there clearly separable clusters of wines? How many clusters did you go with? How the clusters obtained in part (i) are different from or similar to clusters obtained in part (ii), qualitatively?

As seen above, there are clearly 3 clusters of wine we can go ahead with since visually also it seems proper and dendrogram also are clean.

The difference in the clusters obtained via using all chemicals vs 2 Pcs have slight difference in terms of the count of wines included and certain chemical content. We have plotted the same below –

Chem

Chem Membership	Count of Chem Membership	Avg. Alcalinity. of.ash	Avg. Alcohol	Avg. Ash	Avg. Color. intensity	Avg. Flavanoids	Avg. Hue	Avg. Magnesium	Avg. Malic.Acid	Avg. Nonflavanoid.phenols	Avg. OD280. OD315.of. diluted.wi..	Avg. Proanthocyanins	Avg. Proline	Avg. Total. phenols
1	64	17.53	13.67	2.46	5.45	3.01	1.07	106.16	1.97	0.29	3.16	1.91	1,076.05	2.85
2	58	20.21	12.20	2.22	2.90	2.09	1.06	92.55	1.94	0.36	2.86	1.69	501.43	2.26
3	56	21.00	13.06	2.41	6.85	0.85	0.72	99.86	3.17	0.45	1.73	1.13	624.95	1.69

PC

PC Membership	Count of PC Membership	Avg. Alcalinity. of.ash	Avg. Alcohol	Avg. Ash	Avg. Color. intensity	Avg. Flavanoids	Avg. Hue	Avg. Magnesium	Avg. Malic.Acid	Avg. Nonflavanoid.phenols	Avg. OD280. OD315.of. diluted. wines	Avg. Proanthocyanins	Avg. Proline	Avg. Total. phenols
1	61	17.50	13.69	2.47	5.48	3.01	1.06	108.13	2.01	0.29	3.17	1.93	1,101.54	2.85
2	69	19.86	12.28	2.22	3.05	2.03	1.05	93.16	1.90	0.36	2.75	1.58	516.42	2.23
3	48	21.51	13.16	2.44	7.41	0.79	0.68	98.54	3.39	0.46	1.70	1.17	627.50	1.68

As we can see there are minor differences in the clusters formed via both methods. For eg. Cluster 2 via method (i) has 58 wines, whereas (ii) as 69. We can account for this different due to the fact that (ii) incorporates only 55% variance of the data. Similar differences in means of chemical composition are noted.

However, we can safely say that (ii) despite covering only 55% variance of the data, has captured the clusters well.

Q(e) - Could you suggest a subset of the chemical measurements that can separate wines more distinctly? How did you go about choosing that subset? How do the rest of the measurements that were not included while clustering, vary across those clusters?

We can select a subset of the chemical measurement that separate the wine by looking at the eigen vectors which have majority impact on first few PCs.

Eigencevtor Loadings

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
Alcohol	0.144	0.484	0.207		0.266	0.214		0.396	0.509	0.212
Malic.Acid	-0.245	0.225		-0.537		0.537	-0.421			-0.309
Ash		0.316	-0.626	0.214	0.143	0.154	0.149	-0.170	-0.308	
Alcalinity.of.ash	-0.239		-0.612			-0.101	0.287	0.428	0.200	
Magnesium	0.142	0.300	-0.131	0.352	-0.727		-0.323	-0.156	0.271	
Total.phenols	0.395		-0.146	-0.198	0.149			-0.406	0.286	-0.320
Flavanoids	0.423		-0.151	-0.152	0.109			-0.187		-0.163
Nonflavanoid.phenols	-0.299		-0.170	0.203	0.501	-0.259	-0.595	-0.233	0.196	0.216
Proanthocyanins	0.313		-0.149	-0.399	-0.137	-0.534	-0.372	0.368	-0.209	0.134
Color.intensity		0.530	0.137			-0.419	0.228			-0.291
Hue	0.297	-0.279		0.428	0.174	0.106	-0.232	0.437		-0.522
OD280.OD315.of.diluted.wines	0.376	-0.164	-0.166	-0.184	0.101	0.266			0.137	0.524
Proline	0.287	0.365	0.127	0.232	0.158	0.120		0.120	-0.576	0.162

We can also check the correlation of the variables with the PCs and understand the impact.

Correlation Matrix

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Alcohol	0.313093350	0.764257253	0.2493833	0.01711761	0.24539445	0.17105193
Malic.Acid	-0.531884726	0.355431713	-0.1070404	-0.51467982	-0.03252695	0.43000667
Ash	-0.004449362	0.499446109	-0.7530514	0.20531539	0.13211313	0.12373961
Alcalinity.of.ash	-0.519157081	-0.016734916	-0.7360433	-0.05834174	-0.06105952	-0.08076395
Magnesium	0.308022936	0.473476124	-0.1572388	0.33724320	-0.67157727	0.03055463
Total.phenols	0.856136658	0.102774237	-0.1757842	-0.18987451	0.13792594	-0.06738490
Flavanoids	0.917470177	-0.005309113	-0.1811991	-0.14599455	0.10070755	-0.01515560
Nonflavanoid.phenols	-0.647607018	0.045476816	-0.2048724	0.19489072	0.46250110	-0.20714285
Proanthocyanins	0.679921705	0.062103856	-0.1797229	-0.38254807	-0.12641790	-0.42758878
Color.intensity	-0.192235968	0.837489383	0.1651145	-0.06319842	0.07060492	-0.33534860
Hue	0.643662066	-0.441242229	-0.1024817	0.41007505	0.16036834	0.08489588
OD280.OD315.of.diluted.wines	0.816018903	-0.259933849	-0.1996251	-0.17650390	0.09344276	0.21295601
Proline	0.622050797	0.576612723	0.1524154	0.22247039	0.14582396	0.09590437

From both the above, we understand that if we want to capture 55% of the variance (using PC1 & PC2), the chemical subset most important is –

Total Phenol, Flavanoid, OD 280, Alcohol, Proline.

Since they have a high impact of the first 2 PCs.

If we want more variance to be covered, we should incorporate more chemicals further down the columns having high impact.