# Using PCA and Preference Mapping to Investigate Sensory Attribute Variability and Relationships in Cocoa Bean Liquor

## Dataset

The two types of data used in this sensory analysis. Sensory profiling and consumer liking data are both prevalent in sensory and consumer science. This is due to the type of statistical method used for the sensory analysis. Descriptive analysis / sensory proﬁling data can be analysed with PCA with preference mapping together with consumer liking data; with PLSR or PCR together with either product design data or consumer characteristics data.

The sensory attribute data for cocoa beans from ten different farms were processed into liquor for the sensory profiling research. The dataset contains 15 attributes that were assessed by 100 panels. The dataset's attributes are as follows: **" Floral, Liquor, Spicy, Herbal Roasted, Animal-farm, Cocoa, Bitter, Dried fruit, Astringent, Fruity, Acid, Wood, Smoky, Hearthy-mouldy".**

Using a binary coding system, the panels tasted the qualities and recorded their presence or absence. If the attribute was present in the taste, a value of 1 was assigned, and a value of 2 was assigned if it was not.

The consumer liking data was collected through a controlled experiment involving 100 panels and ten different cocoa bean liquors. Each consumer in the panel rated their liking of the liquor on a hedonic scale. The evaluation was based on the assessment of the 15 sensory attributes associated with the liquors. The primary objective of this experiment was to investigate the potential influence of specific sensory attributes on consumer liking of the cocoa bean liquor. Thirthy panels were chosen at random for this analysis.

The sensory profiling data was collected to explore how the samples from different farms relate to each other based on their sensory profiles and to analyse the predominant attributes sensed by the panellists. The final dataset used for the sensory profiling study is a summary count of each farm's attributes. This summary count is likely to indicate the frequency or number of times each attribute was perceived by the panels for each cocoa bean sample from a particular farm.

The dataset will aid in exploring the sensory features of cocoa beans from various farms and understanding the relationships between the sensory attributes and the provenance of the samples. The research intends to identify the most often experienced attributes and investigate how these traits change between farms by assessing summary count data, providing insights into the sensory profile of the cocoa beans and potential variances based on farm origin.

The table below contains the final sensory profile data and consumer liking data used in the analysis.

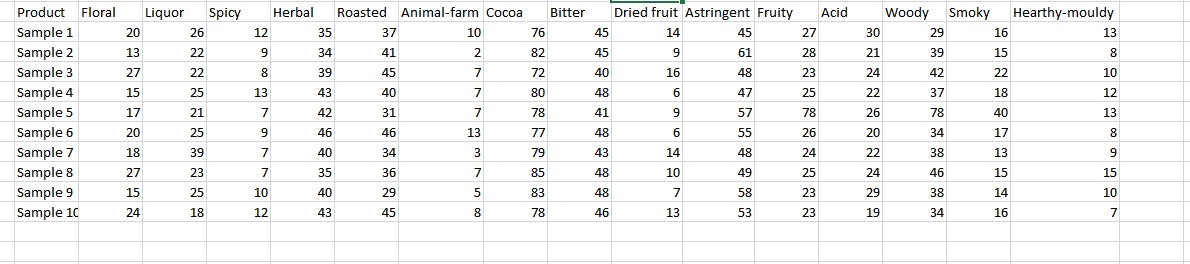


Figure 1: Sensory Profile Data

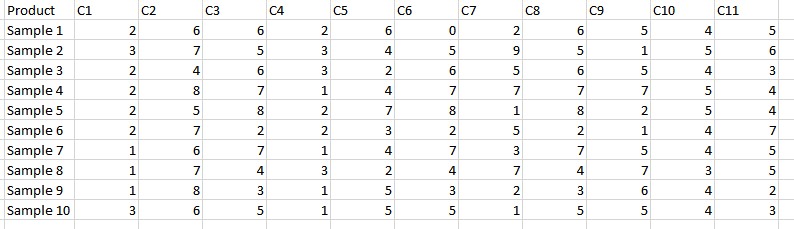


Figure 2: Consumer Liking Data

## Principal Component Analysis (PCA)

In this research, PCA is used as the main statistical method, PCA, or Principal Component Analysis, is a statistical method used in the analysis of sensory attribute data for cocoa beans. PCA is a dimensionality reduction technique which aims to capture the most important information in a high-dimensional dataset by projecting it onto a lower-dimensional space. It determines the primary components, which are linear combinations of the original qualities that account for the greatest amount of variation in the data.

PCA can be applied to a sensory attribute data for cocoa beans from several farms in the sensory profiling analysis. The goal is to acquire insight into the links between the samples and determine the predominant attributes that contribute to the variation observed in the data.

According to Næs et al. (2010), "For PCA in general, many different types of standardization are used, but here we confine ourselves to the most used, namely division by standard deviation." Therefore, we standardize the data to prevent attributes with larger scales or variances from dominating the analysis. This step ensures that all attributes have a comparable impact on the results.

The original data is projected onto the chosen principal components, yielding a new collection of variables known as the scores. These scores indicate the coordinates of the samples in the reduced-dimensional space created by the principal components.

To determine which attributes, have the greatest impact on each component, the loadings of the original attributes on the primary components are analysed. The predominant attributes that contribute to the observed change in the sensory attribute data can be found by examining the loadings.

The explained variance ratios for each primary component are displayed using a scree plot. Based on the large explained variance, this figure aids in deciding how many components to keep. In the reduced-dimensional space provided by the principal components, biplots or scatter plots can also be used to show the correlations between the cocoa bean samples. We may then see how the samples relate to one another depending on their sensory characteristics.

## Preference Mapping

Preference Mapping is a statistical method for understanding the relationships between product attributes and consumer preferences that is used in the analysis of consumer liking data and sensory profiling data Preference Mapping can be used to analysed the sensory attribute data of cocoa beans from 10 different farms, which were processed into liquor and evaluated by 100 panels. This can help identify the predominant attributes perceived by the panellists and examine how the samples from different farms relate to each other based on these attributes.

Preference mapping visualises individual differences between consumers and their preference for products with certain sensory attributes. The preference mapping model is actually a multivariate regression model that consists of an X and Y matrix and that attempts to ﬁnd components that describe common variation between the two. Depending on whether the consumer liking data or descriptive analysis / sensory proﬁling data is chosen to be the X matrix, one speaks of internal or external preference mapping [Næs et al., 2010], respectively. Furthermore, for the computation of the components one can choose between partial least square regression (PLSR) and principal component regression (PCR).

Correspondence Analysis, a common statistical technique in preference mapping, is used to study the relationship between sensory attributes and samples. Correspondence Analysis generates a perceptual map, which visualizes the location of the samples and attributes in a reduced-dimensional space. The map depicts the proximity of samples and attributes, indicating similarities and differences as perceived by the panellists.

The perceptual map is interpreted to gain insights into the dominant attributes and their relationships to the samples and determine clusters or groups of samples with similar sensory profiles. Analysing attribute positioning allows us to better understand their impact on panellists’ perceptions of cocoa bean samples. We identify the key characteristics that most significantly contribute to the differences between samples.

Preference mapping will focus on understanding the sensory attributes perceived by panellists and investigating how cocoa bean samples from different farms relate to each other based on these attributes. In addition, we hope to provide insights into the sensory characteristics of cocoa beans and their variations based on farm origin, allowing for more informed decisions in product development and quality control.

## Descriptive Statistics for Consumer liking data.

The purpose of this section is to provide visualisation and descriptive analysis on the consumer liking data. In this section various plots will be analysed to give us more insight of the consumer liking data. A box plot is generated to analyse the ratings of each sample by the panellists.

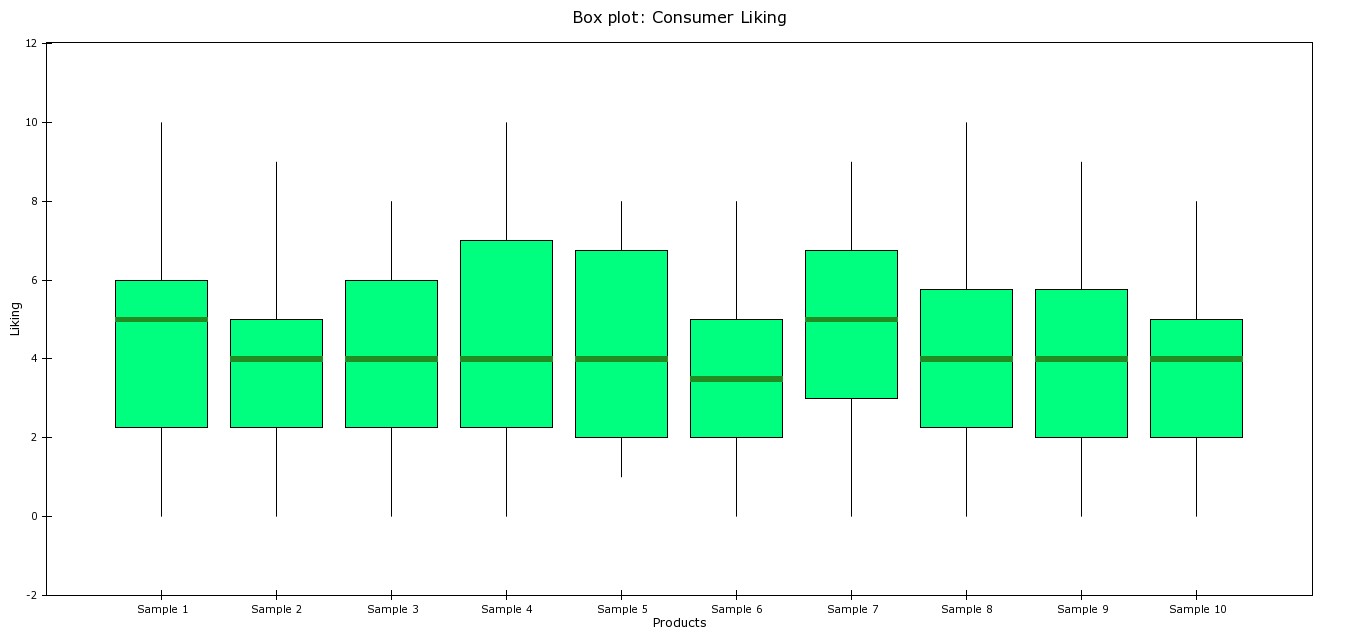
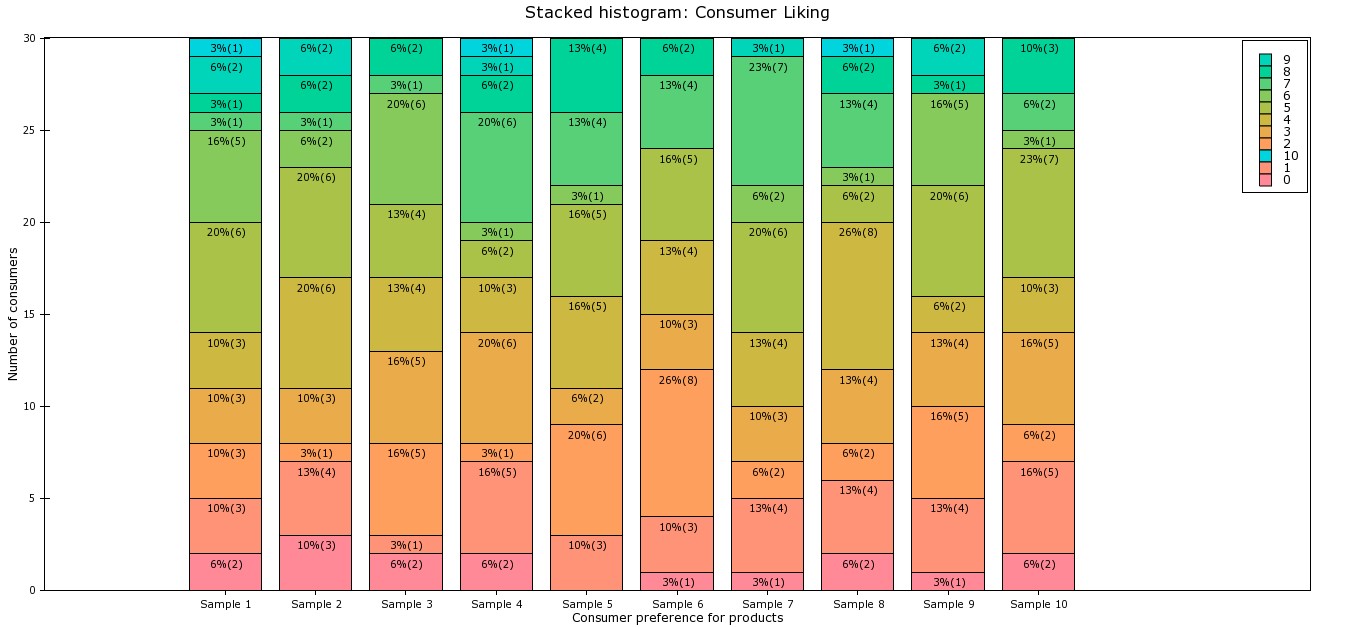


Figure 3:Box Plot of Consumer Liking

Figure 4: Stacked Histogram for Consumer Liking



Summary Highlights:

The box plot in Figure 3 shows the cocoa bean liquor consumer liking data where consumers rated 10 cocoa bean liquor samples, each obtained from 10 different farms, on a hedonic scale from 1 to 9, where 1 represents "don’t like at all" and 9 represents "like very much" based on the sensory attributes. The box plot shows that across all consumers, each cocoa bean liquor received the highest (9) and lowest (1) rating by at least one consumer. This is visualized by the vertical lines that extend from 1 to 9 for each product. The green boxes for each line visualize the distribution of the ratings between the 25th and 75th percentiles. The dark green line across the green boxes shows the median rating for that product. From the boxplot, we observe that sample 1 and sample 7 has a higher median rating compared to the rest of the samples; this implies that the cocoa bean liquor in the first and seventh farm has better consumer liking than the rest of the samples. This claim is solidified as we observe the whiskers for each sample. The cocoa bean liquor in the first and seventh farm had a wider interquartile range, implying a more diverse range of opinions among consumers compared to the other samples.

The stacked histogram in Figure 4 provides another and richer way of visualising consumer liking data. Along the horizontal axis again the cocoa bean liquor samples are shown, while the vertical axis displays either the number of consumers. With the stacked histogram each bar represents one product and each colour in the bar represents a certain rating of the samples. For sample 7, one can see 65% of the total number of consumers rated this sample high. Same with sample 1 as 61% of the total number of consumers rated this sample high. This also implies consumers gave a more positive rating to the cocoa bean liquor in the first and seventh farm as compared to other samples.

## Principal Component Analysis for Sensory Profile Data.

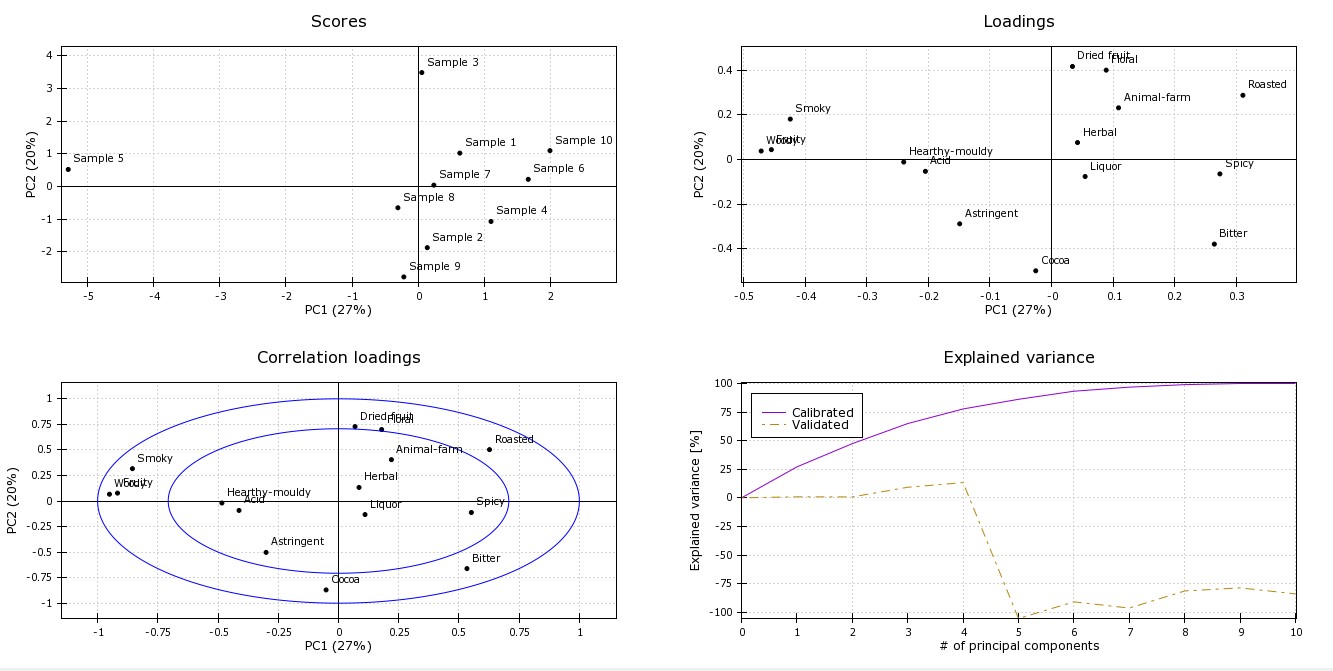
This section provides an in-depth analysis of applying principal component analysis on the sensory profile data. Many plots will be displayed and interpreted. Before analysing the sensory profile data with PCA, the data needs to be standardised. Standardization is the process of transforming the original variables or attributes in a dataset to have a common scale and centring them around zero mean. It is an important step in PCA to ensure that variables with different measurement units or scales do not dominate the analysis. Since the sensory profile data consist of the summary count of panel’s judgement of attributes in a sample, data points are of different scales, this may affect the overall performance of the PCA this is because, it goes to the core of how to interpret the sensory attributes and to how the assessors are trained and calibrated to correct this, we apply standardization by divided the observations for each variable by its standard deviation. This will set the observations on the same scale normally from 0 – 1. After standardization we apply PCA on the data by reducing the dimensions of the data will retaining most variability (import information) in the data. Most PCA plots used in this analysis are in two components, this is because, “the first component describes the most of the variability, the second is the next in the order etc. A consequence of the criterion used is that variables or variable groups with large variance will have a stronger impact on the solution than the rest. Usually, one extracts only a few components treating the rest of the variability as noise.” Næs et al. (2010). The summary plots after PCA are shown below;

Figure 5: Overview plot for the PCA model based on the cocoa bean liquor sensory Profiling.

### PCA scores plot

In sensory profiling, multiple attributes are evaluated for each sample, including floral, liquor, spicy, herbal, roasted, animal-farm, cocoa, bitter, dried fruit, astringent, fruity, acid, woody, smoky, and earthy-mouldy. In a high-dimensional space, these attributes can be represented as variables or dimensions.

PCA reduces data dimensionality by transforming the original variables into a smaller set of uncorrelated variables known as principal components (PCs). Each PC captures a portion of the variation present in the original data. The first PC explains the most significant variation, followed by the second PC, and so on.

Using PCA, cocoa bean liquor sensory profiling data is analysed using four plots. Among these plots are the PCA score plot, the loading plot, the correlation loading plot, and the explained variance plot. Based on these plots, one can identify the predominant attributes that influence sensory attribute variation. First, we analyse the score plot.

Using PCA score plots, samples are visualized in two- or three-dimensional space according to their principal component scores. A point represents each sample on the plot, and their proximity indicates similarity in their sensory profiles.

By analysing the sensory attributes of different samples, one can understand the relationships between them. In this way, it is possible to identify groups or clusters of samples that share similar sensory characteristics. Data patterns, trends, and outliers can be revealed by this plot. Samples of cocoa bean liquor may differ in quality, characteristics, or differences based on these patterns.

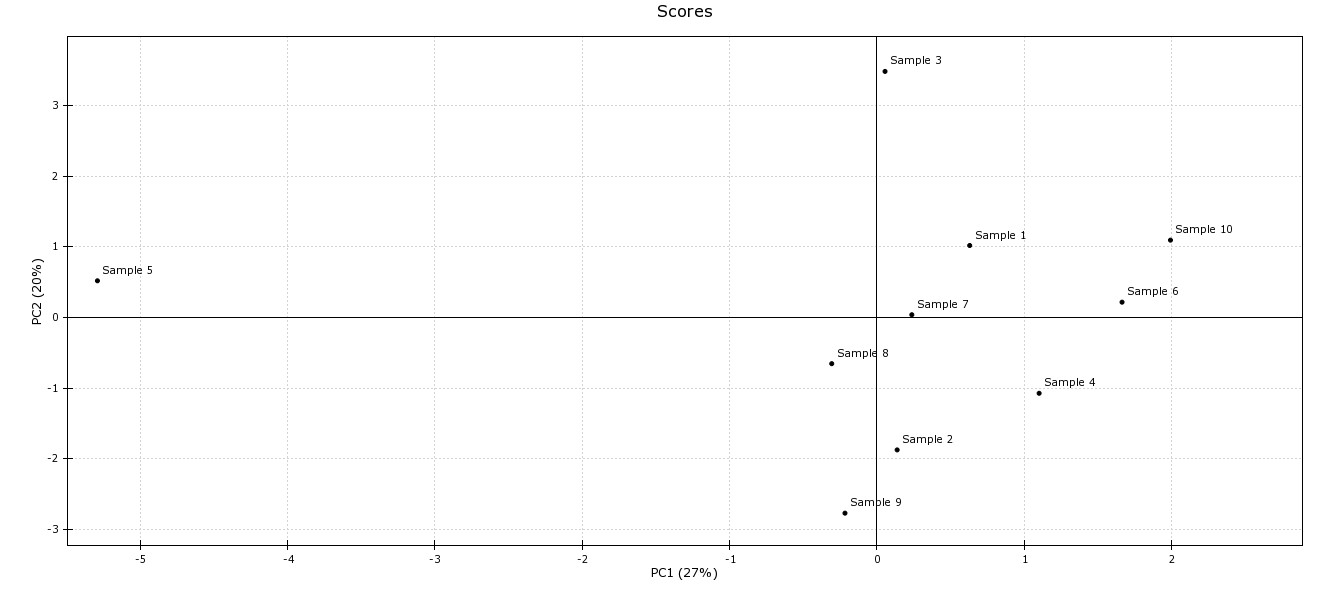
The PCA score plot in analysing cocoa bean liquor sensory profiling data visually represents the relationship and grouping of the samples. This is based on their sensory attributes. 

Figure 6: Score Plot for sensory profile data of the samples

Summary Highlights:

From the score plot, PC1 explains 27% of the total data variance, while PC2 explains 20% of the total variance. Together, these two principal components capture significant variation in the sensory attribute data. The plot is divided into four quadrants based on the positions of the samples. Each quadrant indicates a distinct combination of the samples' PC1 and PC2 scores. Sample 5, placed in the upper left quadrant, is an outlier that stands out from the rest of the samples. It possesses specific sensory features when compared to the other samples. Compared to the rest of the samples, it has distinct sensory characteristics. Sample 3 is located in the upper right quadrant, close to the y-axis, indicating that it has high scores on PC2. Sample 1 and Sample 7 are close to each other in this quadrant, implying some similarity in their sensory profiles. Similarly, Sample 6 and Sample 10 are also close to each other, suggesting shared sensory attributes. Sample 8 is closer to the origin in the lower left quadrant, indicating lower scores on both PC1 and PC2. Sample 9 is closer to the y-axis, indicating higher scores on PC2. Sample 2 is closer to the y-axis in the lower right quadrant, indicating higher scores on PC2. Sample 4 also lies in this quadrant, indicating a different sensory profile than the others.

In summary, the PCA score plot reveals relationships and similarities between cocoa bean samples based on sensory attributes. Samples' positioning in different quadrants reflects differences in their sensory profiles. The plot reveals potential groupings or clusters of samples with comparable sensory characteristics, such as Sample 1's proximity to Sample 7, and Sample 6 and 10. The unusual location of Samples 5and 3 demonstrates that they have distinct sensory profiles when compared to the other samples. However, samples in the same quadrant that are reasonably close together show comparable patterns of variation in their sensory attributes. Therefore, samples 1, 7 and 8, which are in the same quadrant and very close together, have similar sensory characteristics or profiles. The same is true for samples 7 and 8 and samples 2 and 9. They are in different quadrants but close together. The plot aids in understanding the underlying structure and trends in the sensory attribute data and can be used to guide further research and interpretation of the sensory profiles of cocoa bean samples.

### PCA Loadings

Using the PCA loadings plot, cocoa bean liquor sensory profiling data can be analysed to determine how each sensory attribute contributes to the principal components. As a result, it is easier to understand which attributes contribute most to observed patterns and variations.

In this analysis, the PCA loading plot identifies the sensory attributes with the strongest associations to the principal components. In the plot, the vectors or arrows show the correlations between the sensory attributes and the principal components. As indicated by the length and direction of the of the arrows, the correlation is strong and in a particular direction.

In sensory profiling, the loadings plot can be used to identify key sensory attributes that contribute to variation. As a result, further analysis can be conducted, such as identifying clusters of attributes, studying their relationships, or selecting certain attributes for further study or product improvement.

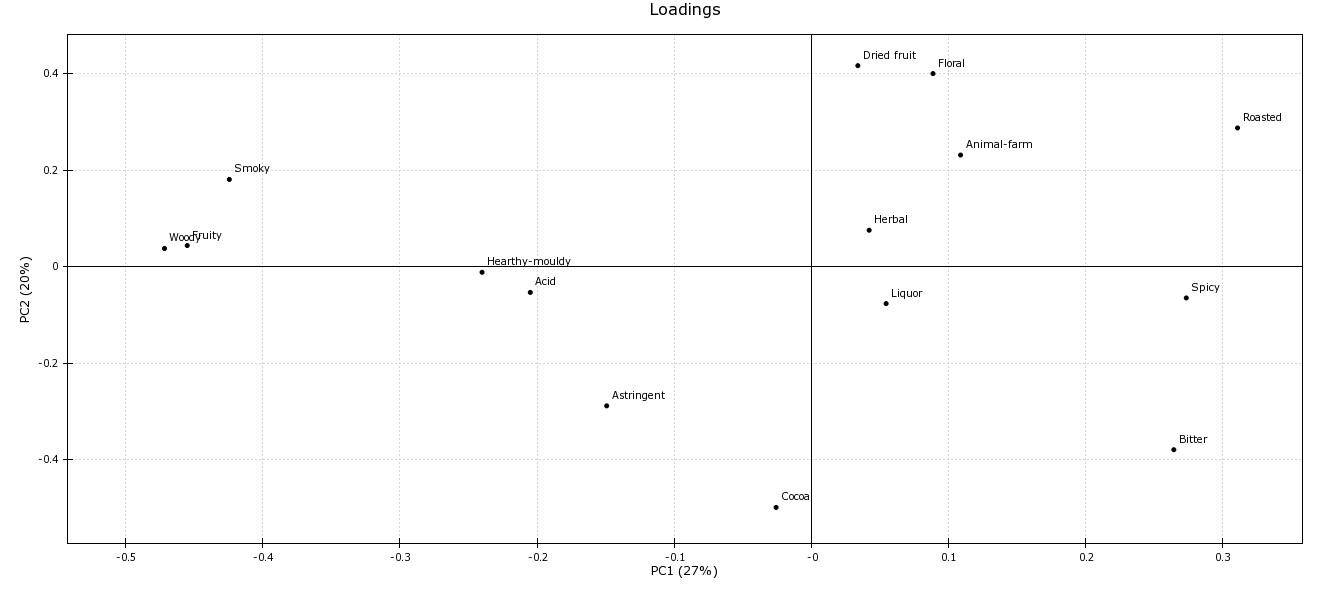


Figure 7: PCA loadings plot for cocoa bean liquor sensory profiling

Summary Highlights:

The attributes with the highest positive loadings in PC-1 are "Roasted," "Bitter," and "Woody," exhibiting a strong positive correlation between these attributes and the first principal component. "Fruity," "Smoky," and "Hearthy-mouldy", on the other hand, have the highest negative loadings in PC-1, indicating a negative association with this component. PC-1 accounts for 27% of total data variance.

In PC-2, the attribute with the highest positive loading is "Floral," which is followed by "Dried fruit" and "Cocoa." This implies that these characteristics have a considerable impact on the second main component. PC-2 accounts for 20% of the overall variation.

The attribute with the highest positive loading in PC-3 is "Acid," whereas "Herbal" has highest negative loading. This implies that PC-3 identifies acidity and herbal variation. PC-3 accounts for 11% of the overall variation.

In terms of PC 4 to PC 10, these components explain lower amounts of variance in the data. Specific attributes with higher loadings in these components may be associated with specific sensory characteristics. For example, "Fruity" has a high negative loading in PC-4, indicating that it plays a role in distinguishing samples based on fruity notes.

More information about the samples can be obtained by superimposing the PCA scores and loadings plot. Both the smoky, woody, and fruity qualities, as well as Sample 5, are found in the upper left quadrant of both the loadings and the score plots. This suggests that sample 5 has a high smoky, woody, and fruity value. In contrast, woody has a strong positive influence for sample 5, while smoky and fruity have a strong negative influence for sample 5.

Hearty-mouldy, acidic, astringent and cocoa, along with samples 8 and 9, are also found in the lower left quadrant of both the load plot and the score plot.

As a result, samples 8 and 9 have a high value for hearthy-mouldy, acid, astringent, and cocoa attributes. Cocoa, on the other hand, has a high positive impact on sample 9 and acid has a strong positive impact on sample 8, whereas heathy-mouldy and astringent have a lesser value for assessing these samples.

Dried fruits, floral, and roasted attributes have a greater value for determining sample 3 and 10. Samples 6, 7, and 10 had a lower herbal value. In the upper right quadrant, Animal Farm had a substantially lesser influence.

Based on this analysis, we can conclude that the attributes: Bitter, Roasted, Smoky, Woody, Cocoa, Floral, Dried fruit, Herbal, Sour, Hearty-mouldy and Animal-farm have the greatest impact on the sensory profiling of cocoa bean liquor.

### PCA Correlations Loadings Plot

PCA correlation loadings, as illustrated in Figure 8, are another technique to visualize the factors' contributions to total variance in the data. PCA correlation loadings indicate how systematic a variable's variance is in relation to the computed PCs. They also show how much variance was contributed by the variable (as shown in the PCA scores plot). Correlation loading, in further detail, assesses the relationship between the original data of a certain variable and the scores of a specific PC. In this way, regardless of the total variance it contributes, one may assess to what extent the variation in a certain variable is systematic or noisy. The two rings in Figure 8's correlation loading plot represent different degrees of explained variance for the attributes under consideration. The outside ring represents 100% of the variance explained, whereas the inner ring represents 50% of the variance explained.

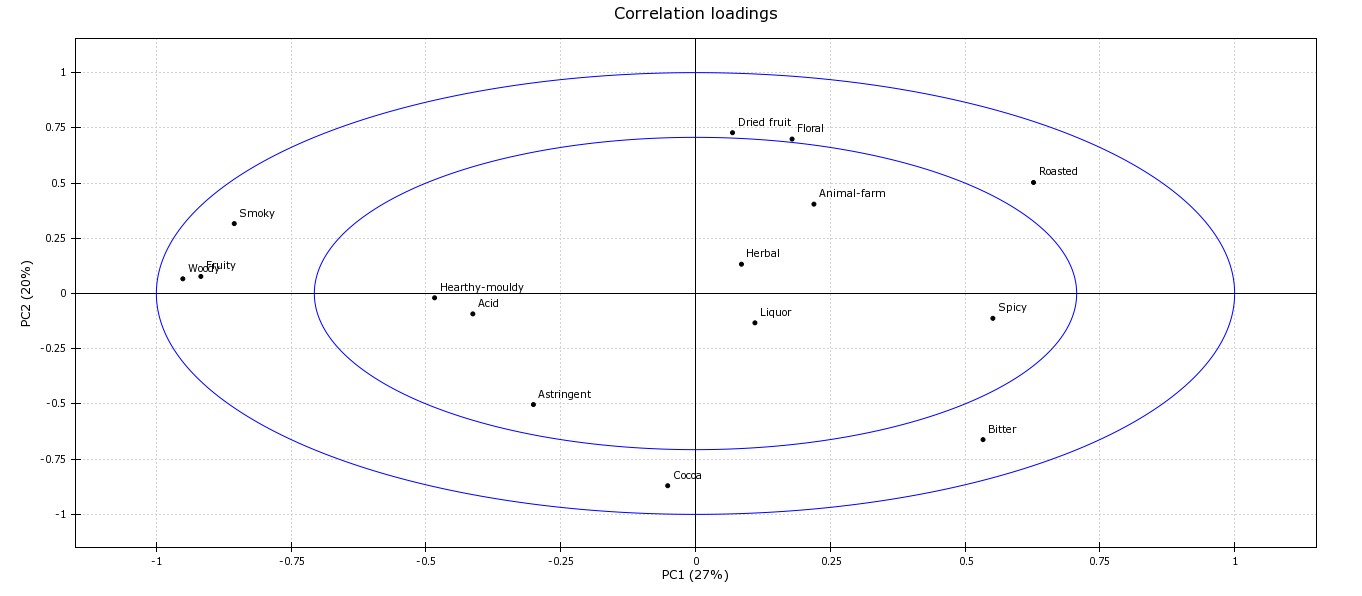


Figure 8: PCA correlation loadings for the cocoa bean liquor sensory proﬁling data.

Summary Highlights:

From the correlation loadings plot, attributes; Herbal, Liquor, Spicy, Astringent, Acid, Hearthy-mouldy and Animal farms. These attributes are located just inside the inner ring, which indicates that just under 50% of the variation of this variable is explained by PC1 and PC2. Remember that that PC1 and PC2 together explain 47% of the total variance in the data and that the remaining higher PCs provide very little or no information. This means that not much more than 50% of the variance of these attributes will be explained by the higher PC’s and that the remaining variance of that variable is likely to be noise. Attributes; Woody, Fruity, Smoky, Dried Fruit, Floral, Roasted, Bitter and Cocoa are very close to the outer ring, meaning that almost 100% of its variation is explained by PC1 and PC2, thus indicating that its variance is very systematic. In this way it is possible to see that the variation of attributes: Woody, Fruity, Smoky, Dried Fruit, Floral, Roasted, Bitter and Cocoa are much more systematic with the variance described by PC1 and PC2 than that of attributes: Herbal, Liquor, Spicy, Astringent, Acid, Hearthy-mouldy and Animal farms.

It is also pertinent to note that the attributes Herbal, Liquor, Spicy, Astringent, Sour, Hearty Mouldy and Animal Farms can be found within the inner ring. The attributes Woody, Fruity, Smoky, Dried Fruit, Floral, Toasted, Bitter and Cocoa are located within the outer ring. This spatial arrangement implies that the attributes in the inner ring are more correlated or similar to those in the outer ring. This means that the sensory profiles of the attributes in the inner ring are more closely related. Similarly, although they may have different patterns or correlations than the attributes in the inner ring, the attributes in the outer ring are more correlated with each other.

Considering this, PCA correlation loadings are a useful complement to the PCA loadings for better understanding of how variables contribute to the total variance in the data.

### PCA Explained Variances.

Figure 9 shows the cumulative calibrated and validated explained variances of the same data. It can be seen easily that with only two PC’s almost all of the variance in the data is explained by the model 47%. Full cross-validation (also known as leave-one-out) was applied to the data for model validation, both scores did not perform well.

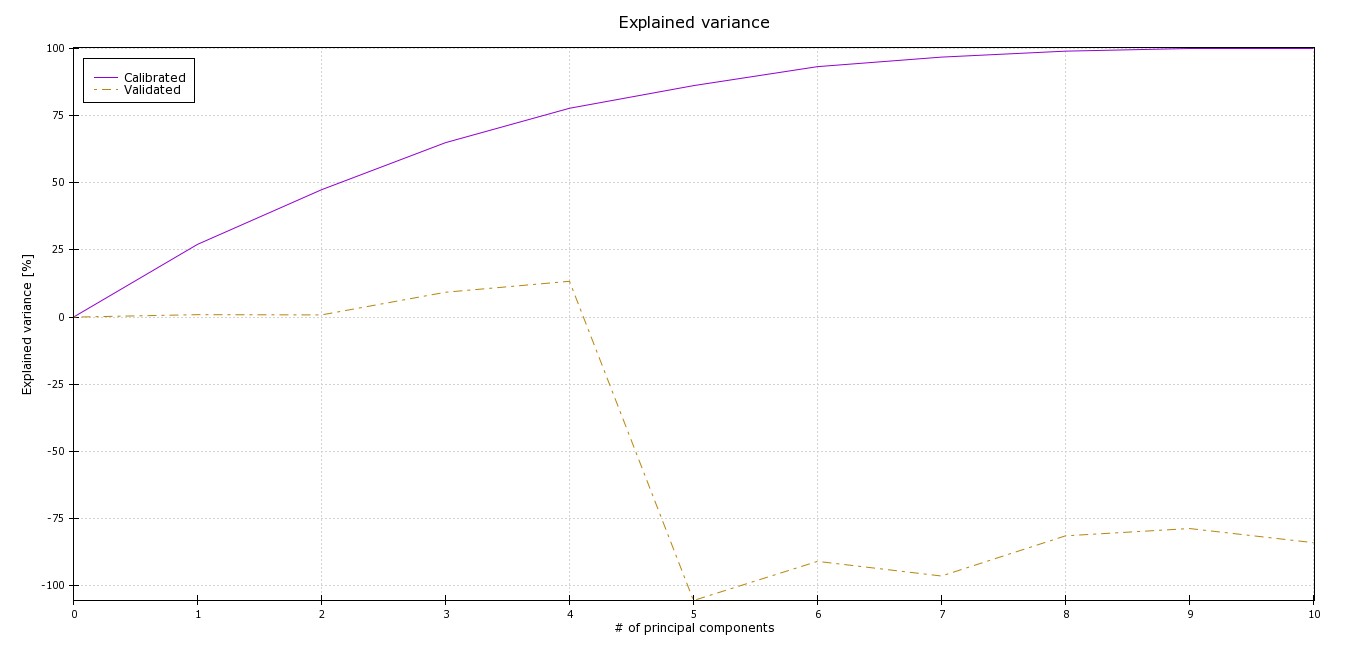


Figure 9: Cumulative calibrated and validated explained variance of the PCA model for the cocoa bean liquor descriptive analysis

## Preference Mapping.

Preference mapping is applied simultaneously to consumer liking data and descriptive analysis / sensory proﬁling data. Preference mapping aims to uncover the relationships between consumer liking and the sensory characteristics of products, providing valuable insights for product development, quality improvement, and market positioning. The two standard statistical tools applied to build a preference mapping model are partial least squares regression (PLSR) and principal component regression (PCR). When building a preference mapping model, both consumers and the trained sensory panel need to evaluate the same set of products. In each data, the row order of the products needs to be identical otherwise wrong data are linked together and results will lead to incorrect conclusions. In the context of cocoa bean liquor sensory profiling, preference mapping plays a crucial role in understanding the factors that influence consumer liking and how sensory attributes contribute to overall preference. By mapping consumer preferences onto sensory space, preference mapping allows researchers to identify the key sensory attributes that drive consumer liking or disliking. Preference mapping provides visual representations, such as preference maps or biplots, that illustrate the relationships between consumer liking, sensory attributes, and product samples. By examining the position of samples and attributes in the preference map, researchers can identify patterns, clusters, and relationships that exist among the sensory attributes and consumer preferences. This knowledge can be used to understand the sensory profiles preferred by consumers, identify the most influential attributes, and guide product formulation, optimization, and marketing strategies. In this section, we will present the results of preference mapping analysis for the cocoa bean liquor sensory profiling data, discussing the relationships between sensory attributes, consumer liking, and product samples. Through this analysis, we aim to gain a deeper understanding of the sensory characteristics that drive consumer preference for cocoa bean liquors and provide valuable insights for the cocoa industry and product development.

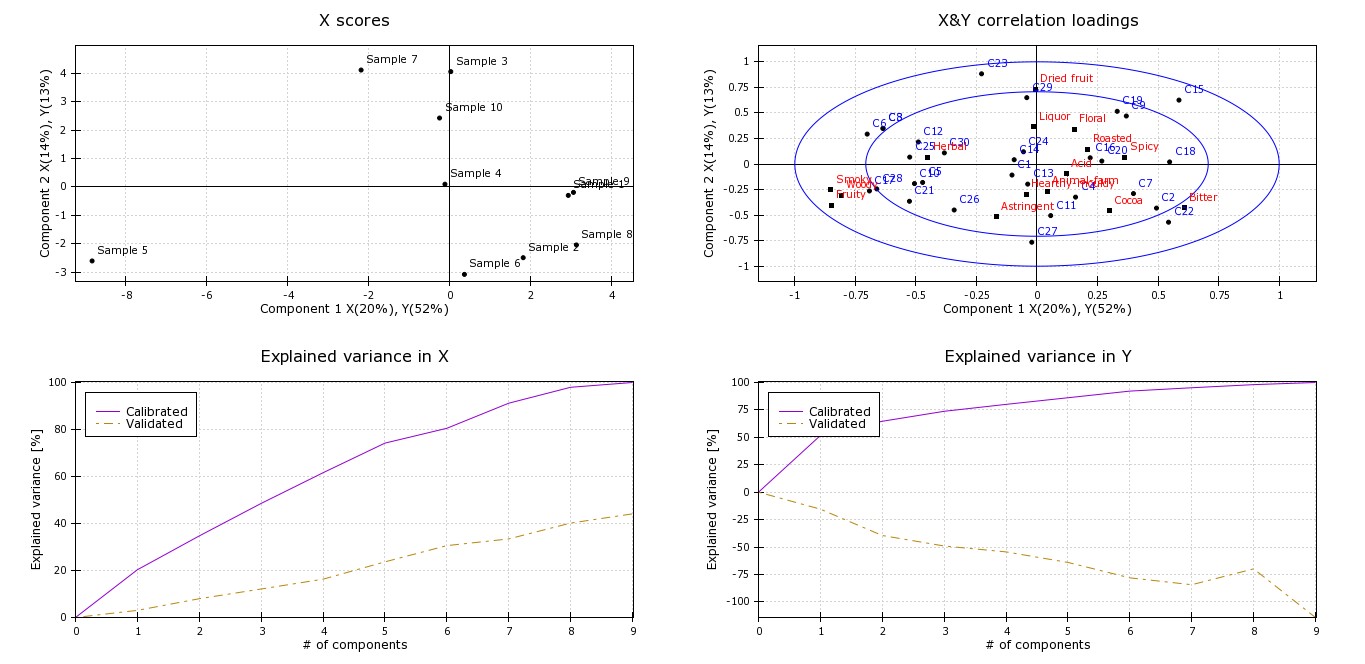


Figure 10: Overview plot for preference mapping model for the cocoa consumer liking data and sensory profiling data.

#### Preference mapping – X scores

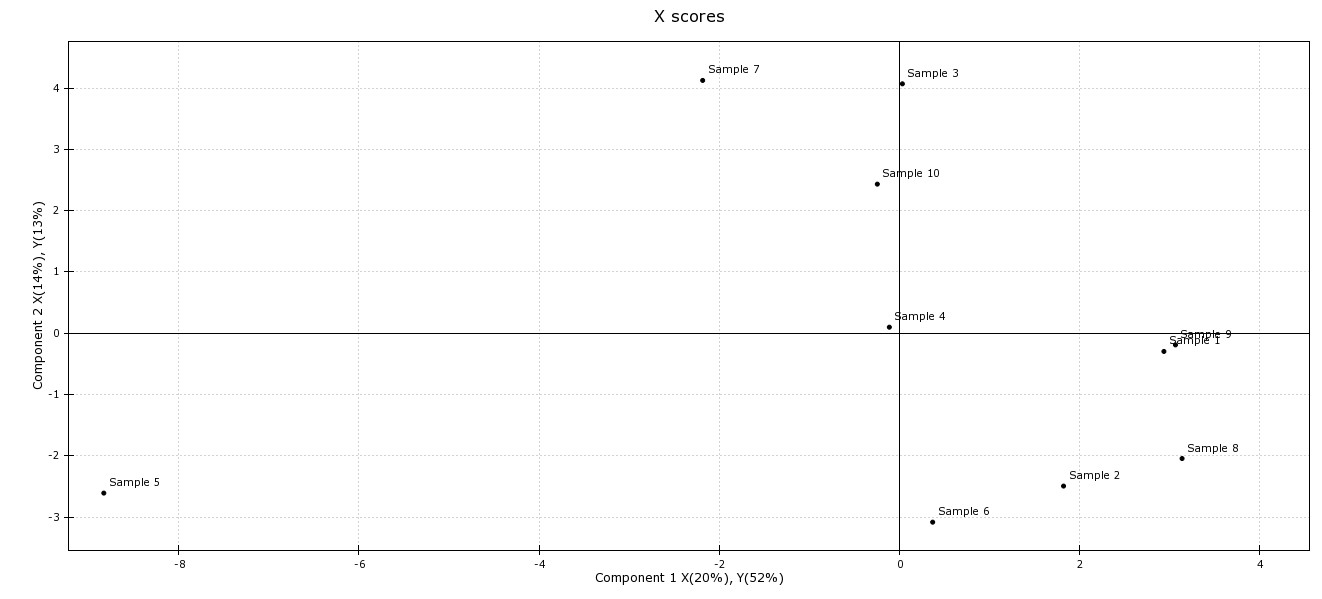
This section presents results computed with the following settings: internal preference mapping (i.e. consumer liking is set to X and descriptive analysis / sensory profiling is set to Y); PLSR; X and Y are not standardised since all variables are calculated at the same scale in the respective matrices. A PCA X score can be used to visually represent similarities and relationships among cocoa liquor sensory profiles. Based on their scores for each principal component, it is possible to distinguish samples with distinct sensory characteristics from those that share similar sensory characteristics. In this way, we will gain a deeper understanding of the sensory characteristics of cocoa liquor and consumers' preferences.

Figure 11: X scores plot in preference mapping for the cocoa consumer liking data and sensory profiling data.

Summary Highlight:

Figure 11 shows the preference mapping model X-scores. This shows how the products relate to each other in the space spanned by the first two components. As with the PCA scores in Figure 6, similar products are close together and dissimilar products are further apart. However, this time the distribution of the products is influenced by the common variance in both the X and Y matrices. In this instance, PLSR was used to compute the preference model. Sample 5 is far away from the other samples in the lower left quadrant, indicating it is an outlier. Sample 5, however, stands out from the rest as it has a higher X-score than the other samples, indicating a significantly higher influence on PC-1 and PC-2 related attributes. This means that sample 5 has very unique attributes that distinguish it from other samples. Samples 1 and 2 have relatively high X-scores, suggesting a relatively strong influence on PC-1 and PC-2 related attributes. Samples 3, 4, 7 and 9 have moderate scores on each PC, indicating a moderate influence on both PC attributes. Sample 6 has a moderate score on PC1 and relatively low scores on PC2, suggesting a moderate influence on both sets of attributes.

#### Preference mapping - X&Y correlation loadings

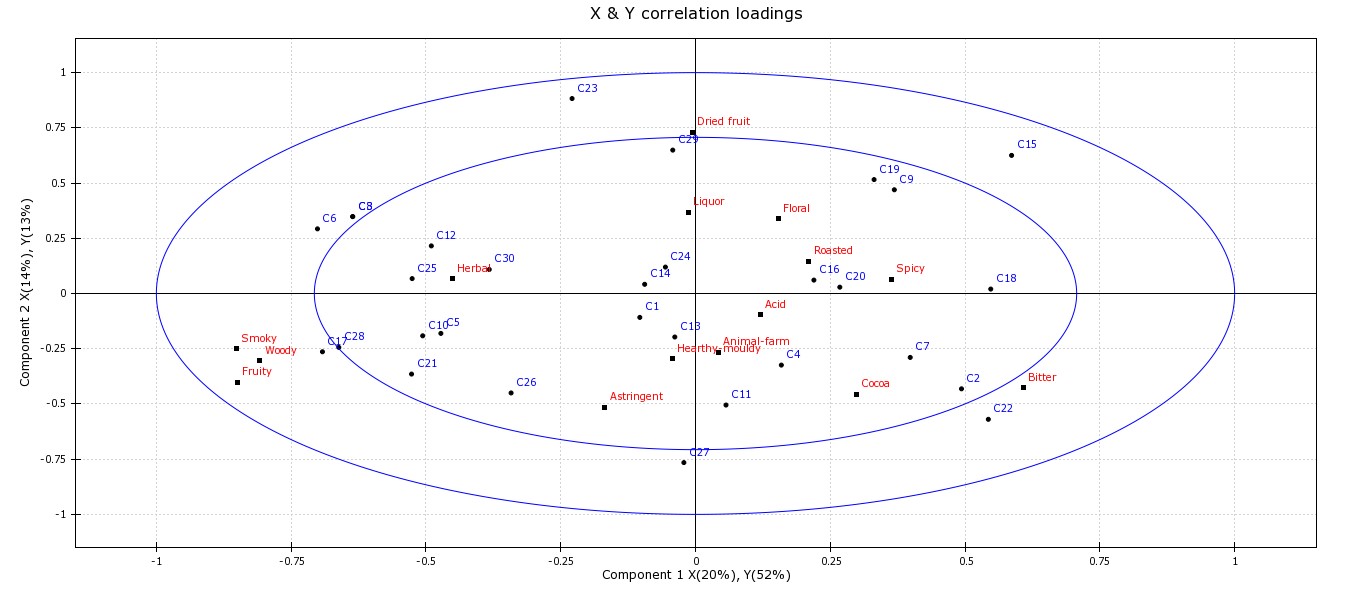


Figure 12: X & Y correlation loadings plot, the actual preference map for the cocoa bean liquor consumer liking data and sensory descriptive analysis / sensory proﬁling data.

Figure 12 illustrates the X & Y correlation loading plot, the actual preference map for cocoa bean liquor consumer liking data. It also shows a descriptive analysis / sensory proﬁling data. The plot below shows the actual preference map. It is used for interpretation and visualization of consumer preferences and sensory attributes that drives consumer liking of a product. In this section, we carefully analyse the sensory attributes in cocoa bean liquor that drive consumer liking of cocoa samples.

Correlation loadings for matrix X are always blue. In this example they start with the letter ’C’ followed by a number that identiﬁes the consumers that participated in the trial. Matrix Y correlation loadings are always red. Specifically, they are the sensory attributes of cocoa samples. What we can conclude from Figure 12 is that many consumers prefer cocoa samples with high intensities of herbal, hearthy mouldy and animal farm attributes (upper left and lower left part of the plot respectively) since a large part of the consumers are in proximity to those attributes. The sensory attributes smoky, woody, fruity and astringent which are all highly correlated, are less preferred although there are a few consumers that prefer high intensities of these sensory attributes. All of them have significant explained variances, since they are located very close to the outer ring which indicates 100% explained variance. Attributes; cocoa, spicy, floral, roasted and acid are also correlated, however to a lesser degree. Moderate consumers prefer high intensities of these attributes. The explained variances for those two attributes are lower. Remember that the inner ring indicates 50% explained variance. Consumers in the inner circle closer to the origin don’t discriminate between the products regarding the variation described by PC1 and PC2.