Multivariate Data Analysis (MVDA) & Visualization

Mpho Mafata

09 October 2024

ECO2WINE

Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

The winemaking process

ECO2WINE

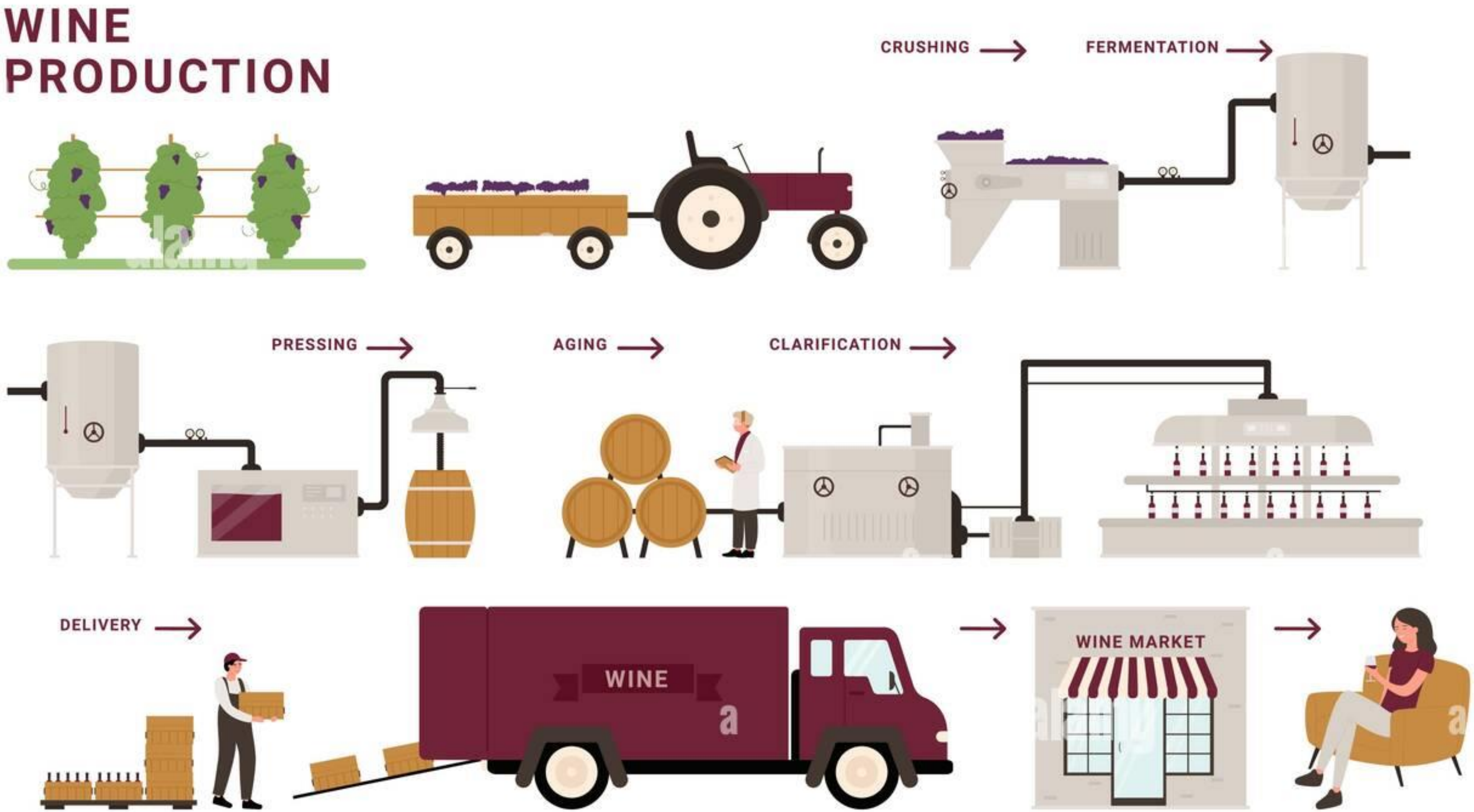


Image source: alamy.com

Univariate to multivariate analysis

observation	Var 1	Var 2
1	Rep 1	Rep 1
1	Rep 2	Rep 2
1	Rep 3	Rep 3
2	Rep 2	Rep 2

Design of experiment (DoE)

- Biological repeats
- Methodological repeats
- Single phenomenon

Observation	Var 1	Var 2	Var 3
1			
2			
3			
4			

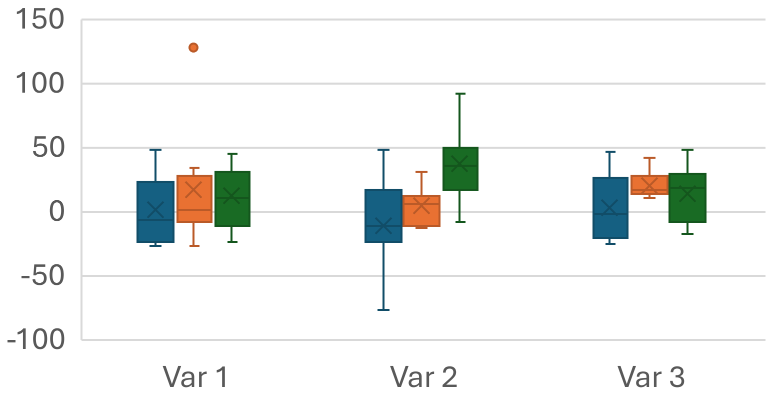
Multiple measurements

- Precision methodology – repeats not needed
- Multiple phenomena

Univariate to multivariate analysis

observation	Var 1	Var 2
1	Rep 1	Rep 1
1	Rep 2	Rep 2
1	Rep 3	Rep 3
2	Rep 2	Rep 2

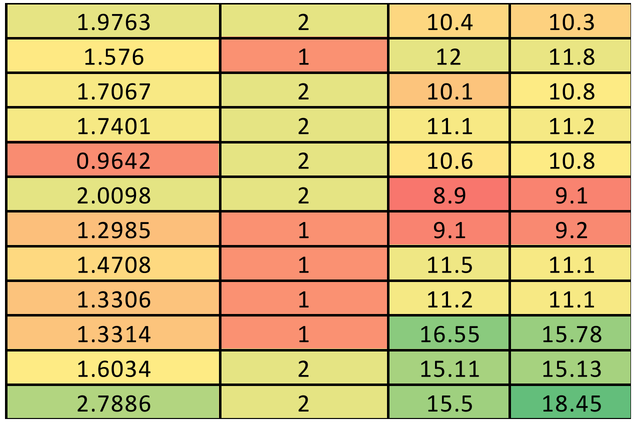
• • •   • • •



Observation	Var 1	Var 2	Var 3
1			
2			
3			
4			

•
•
•

• • •   • • •   • • •   • • •



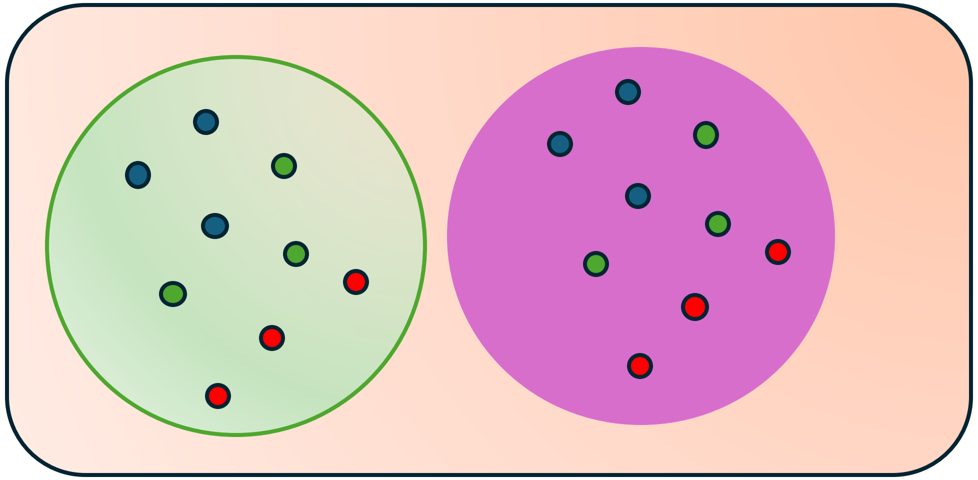
1.9763	2	10.4	10.3
1.576	1	12	11.8
1.7067	2	10.1	10.8
1.7401	2	11.1	11.2
0.9642	2	10.6	10.8
2.0098	2	8.9	9.1
1.2985	1	9.1	9.2
1.4708	1	11.5	11.1
1.3306	1	11.2	11.1
1.3314	1	16.55	15.78
1.6034	2	15.11	15.13
2.7886	2	15.5	18.45

Data wrangling & Caveats in MVDA

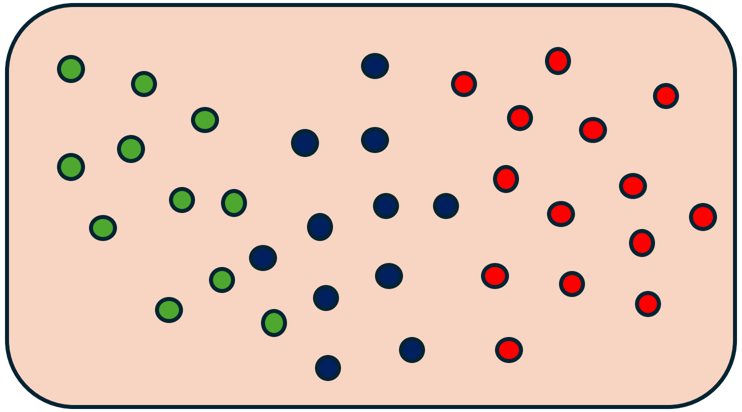
- Aggregation method: sum, counts, etc.
- Treating NaNs – empty vs zeros (consequences?)
- Grouping

Wrangling libraries and techniques:
R - tidyverse
Python – pandas, numpy
SQL - count

Grouping observations
The variance across observational groups of **samples** is greater than between each sample



Grouping variables
The variance across observational groups of **measurements** is greater than between each measurements

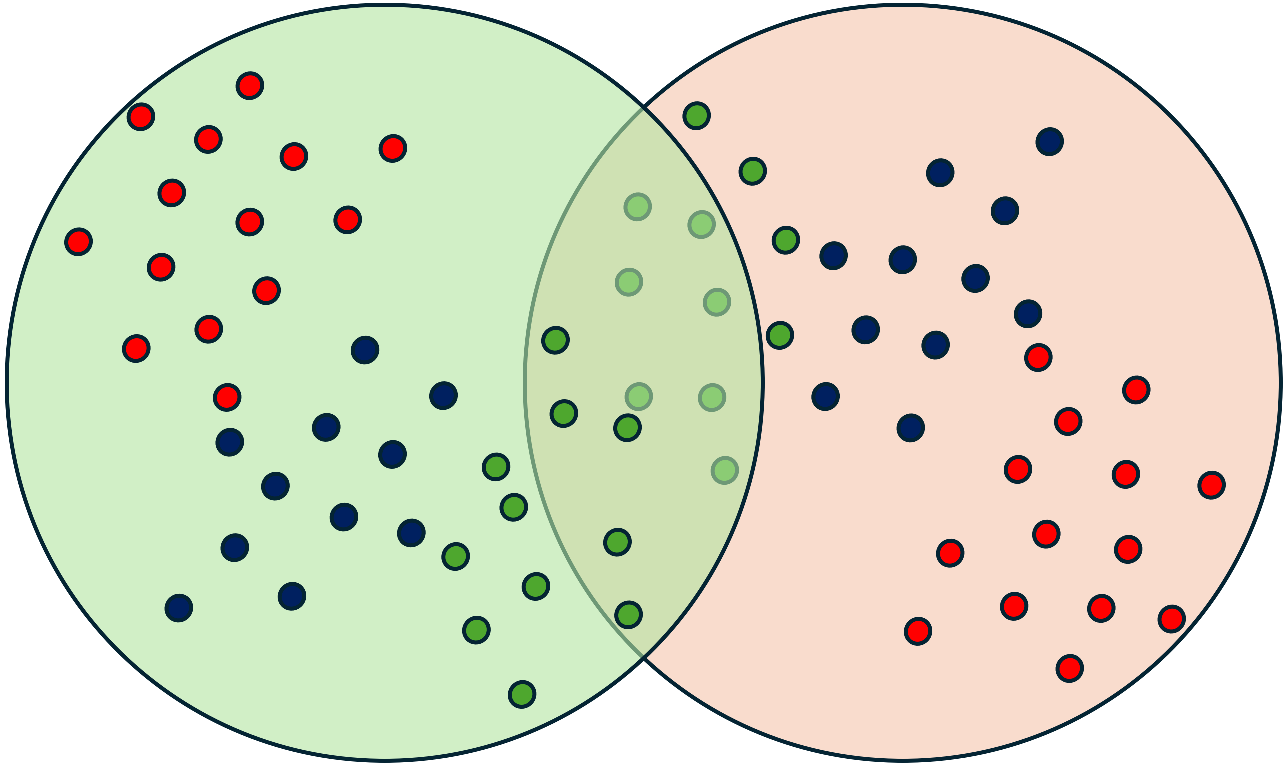


Data wrangling & Caveats in MVDA

- Aggregation method: sum, counts, etc.
- Treating NaNs, NULLs: empty vs zeros (consequences?)
- Grouping

Wrangling libraries and techniques:
R - tidyverse
Python – pandas, numpy
SQL - count

Combined effects



Data wrangling & Caveats in MVDA

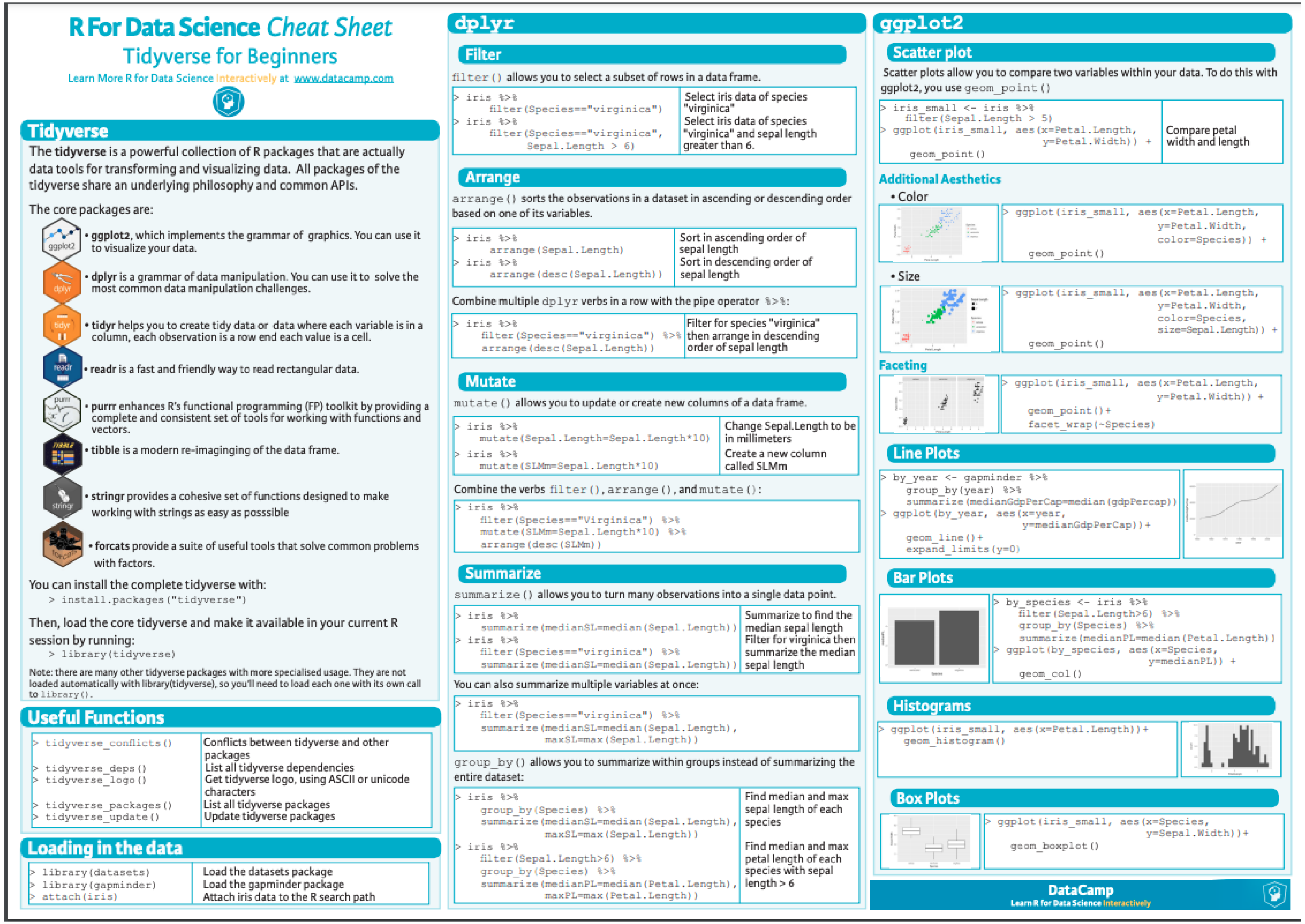
Automate data wrangling and analysis

- Captured format is different from analysis format
- Capture data in the simplest format
- Try to change data in the processing stage instead of in spreadsheets (for reproducibility)

Data wrangling & Caveats in MVDA

Base libraries:

“readxl”
“writexl"



R For Data Science *Cheat Sheet*

Tidyverse for Beginners

Learn More R for Data Science Interactively at www.datacamp.com

Tidyverse

The **tidyverse** is a powerful collection of R packages that are actually data tools for transforming and visualizing data. All packages of the tidyverse share an underlying philosophy and common APIs.

The core packages are:

- **ggplot2**, which implements the grammar of graphics. You can use it to visualize your data.
- **dplyr** is a grammar of data manipulation. You can use it to solve the most common data manipulation challenges.
- **tidyr** helps you to create tidy data or data where each variable is in a column, each observation is a row end each value is a cell.
- **readr** is a fast and friendly way to read rectangular data.
- **purrr** enhances R's functional programming (FP) toolkit by providing a complete and consistent set of tools for working with functions and vectors.
- **tibble** is a modern re-imaginging of the data frame.
- **stringr** provides a cohesive set of functions designed to make working with strings as easy as posssible
- **forcats** provide a suite of useful tools that solve common problems with factors.

You can install the complete tidyverse with:

```
> install.packages("tidyverse")
```

Then, load the core tidyverse and make it available in your current R session by running:

```
> library(tidyverse)
```

Note: there are many other tidyverse packages with more specialised usage. They are not loaded automatically with library(tidyverse), so you'll need to load each one with its own call to library().

Useful Functions

`> tidyverse_conflicts()`	Conflicts between tidyverse and other packages
`> tidyverse_deps()`	List all tidyverse dependencies
`> tidyverse_logo()`	Get tidyverse logo, using ASCII or unicode characters
`> tidyverse_packages()`	List all tidyverse packages
`> tidyverse_update()`	Update tidyverse packages

Loading in the data

`> library(datasets)`	Load the datasets package
`> library(gapminder)`	Load the gapminder package
`> attach(iris)`	Attach iris data to the R search path

dplyr

Filter

filter() allows you to select a subset of rows in a data frame.

```
> iris %>%
    filter(Species=="virginica")
> iris %>%
    filter(Species=="virginica",
           Sepal.Length > 6)
```

Select iris data of species "virginica"

Select iris data of species "virginica" and sepal length greater than 6.

Arrange

arrange() sorts the observations in a dataset in ascending or descending order based on one of its variables.

```
> iris %>%
    arrange(Sepal.Length)
> iris %>%
    arrange(desc(Sepal.Length))
```

Sort in ascending order of sepal length

Sort in descending order of sepal length

Combine multiple dplyr verbs in a row with the pipe operator %>%:

```
> iris %>%
   filter(Species=="virginica") %>%
   arrange(desc(Sepal.Length))
```

Filter for species "virginica" then arrange in descending order of sepal length

Mutate

mutate() allows you to update or create new columns of a data frame.

```
> iris %>%
   mutate(Sepal.Length=Sepal.Length*10)
> iris %>%
   mutate(SLMm=Sepal.Length*10)
```

Change Sepal.Length to be in millimeters

Create a new column called SLMm

Combine the verbs filter(), arrange(), and mutate():

```
> iris %>%
   filter(Species=="Virginica") %>%
   mutate(SLMm=Sepal.Length*10) %>%
   arrange(desc(SLMm))
```

Summarize

summarize() allows you to turn many observations into a single data point.

```
> iris %>%
   summarize(medianSL=median(Sepal.Length))
> iris %>%
   filter(Species=="virginica") %>%
   summarize(medianSL=median(Sepal.Length))
```

Summarize to find the median sepal length

Filter for virginica then summarize the median sepal length

You can also summarize multiple variables at once:

```
> iris %>%
   filter(Species=="virginica") %>%
   summarize(medianSL=median(Sepal.Length),
             maxSL=max(Sepal.Length))
```

group_by() allows you to summarize within groups instead of summarizing the entire dataset:

```
> iris %>%
   group_by(Species) %>%
   summarize(medianSL=median(Sepal.Length),
             maxSL=max(Sepal.Length))
> iris %>%
   filter(Sepal.Length>6) %>%
   group_by(Species) %>%
   summarize(medianPL=median(Petal.Length),
             maxPL=max(Petal.Length))
```

Find median and max sepal length of each species

Find median and max petal length of each species with sepal length > 6

ggplot2

Scatter plot

Scatter plots allow you to compare two variables within your data. To do this with ggplot2, you use geom_point()

```
> iris_small <- iris %>%
    filter(Sepal.Length > 5)
> ggplot(iris_small, aes(x=Petal.Length,
                         y=Petal.Width)) +
    geom_point()
```

Compare petal width and length

Additional Aesthetics

- Color

```
> ggplot(iris_small, aes(x=Petal.Length,
                         y=Petal.Width,
                         color=Species)) +
    geom_point()
```

- Size

```
> ggplot(iris_small, aes(x=Petal.Length,
                         y=Petal.Width,
                         color=Species,
                         size=Sepal.Length)) +
    geom_point()
```

Faceting

```
> ggplot(iris_small, aes(x=Petal.Length,
                         y=Petal.Width)) +
    geom_point()+
    facet_wrap(~Species)
```

Line Plots

```
> by_year <- gapminder %>%
    group_by(year) %>%
    summarize(medianGdpPerCap=median(gdpPercap))
> ggplot(by_year, aes(x=year,
                      y=medianGdpPerCap))+
    geom_line()+
    expand_limits(y=0)
```

Bar Plots

```
> by_species <- iris %>%
    filter(Sepal.Length>6) %>%
    group_by(Species) %>%
    summarize(medianPL=median(Petal.Length))
> ggplot(by_species, aes(x=Species,
                         y=medianPL)) +
    geom_col()
```

Histograms

```
> ggplot(iris_small, aes(x=Petal.Length))+
    geom_histogram()
```

Box Plots

```
> ggplot(iris_small, aes(x=Species,
                         y=Sepal.Width))+
    geom_boxplot()
```

DataCamp
Learn R for Data Science Interactively

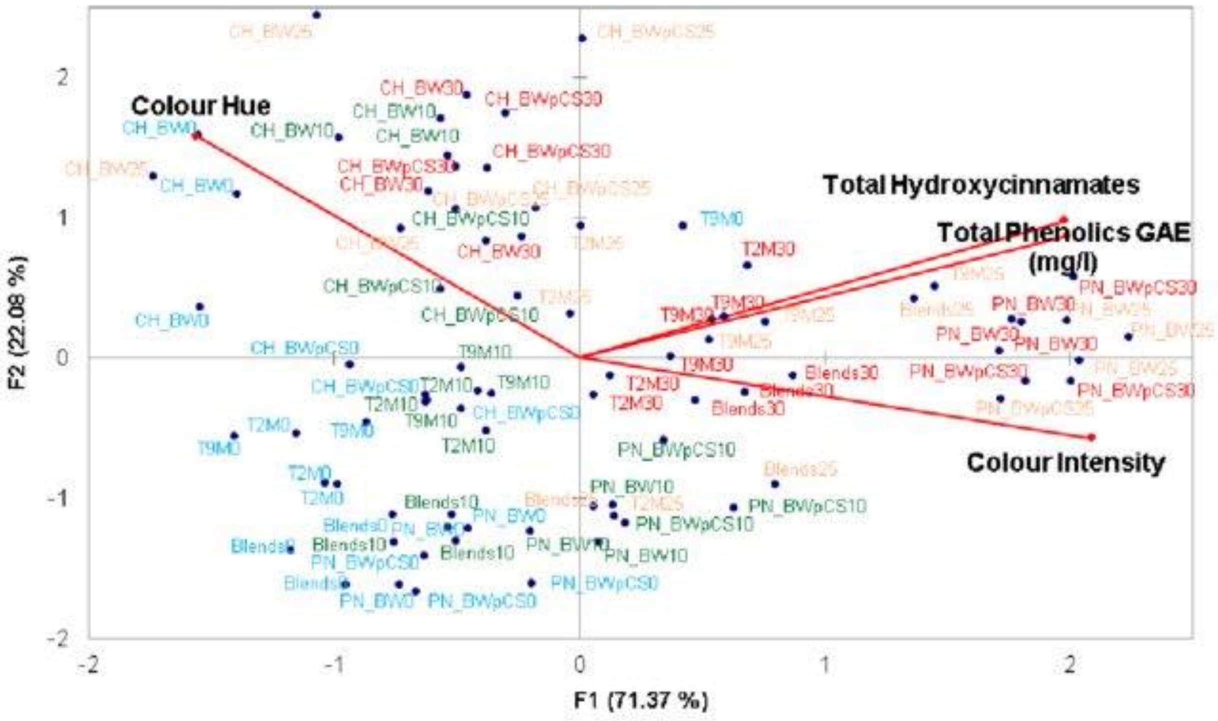
Common MVDA in wine sciences

Dimension reduction

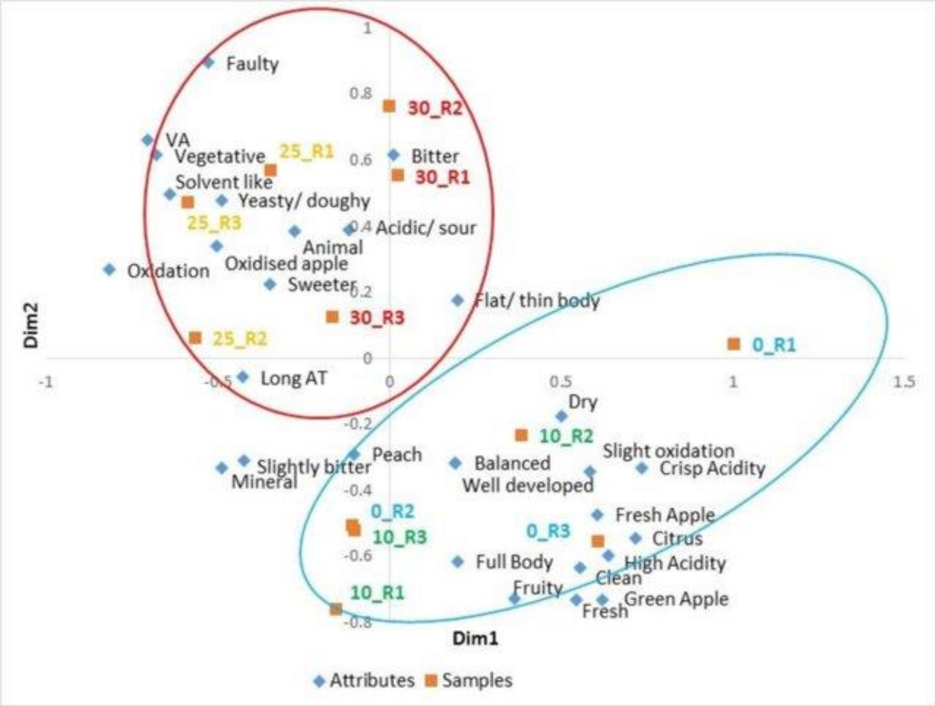
Some come with a common visualization, some do not. Ultimately, they are just **mathematical calculations**, applied according to the type of data (e.g., Categorical, continuous, discreet, nominal, sparse). You get to represent the results in a communicative way that elucidates insight.

1 Orthogonal decomposition

PCA



MDS



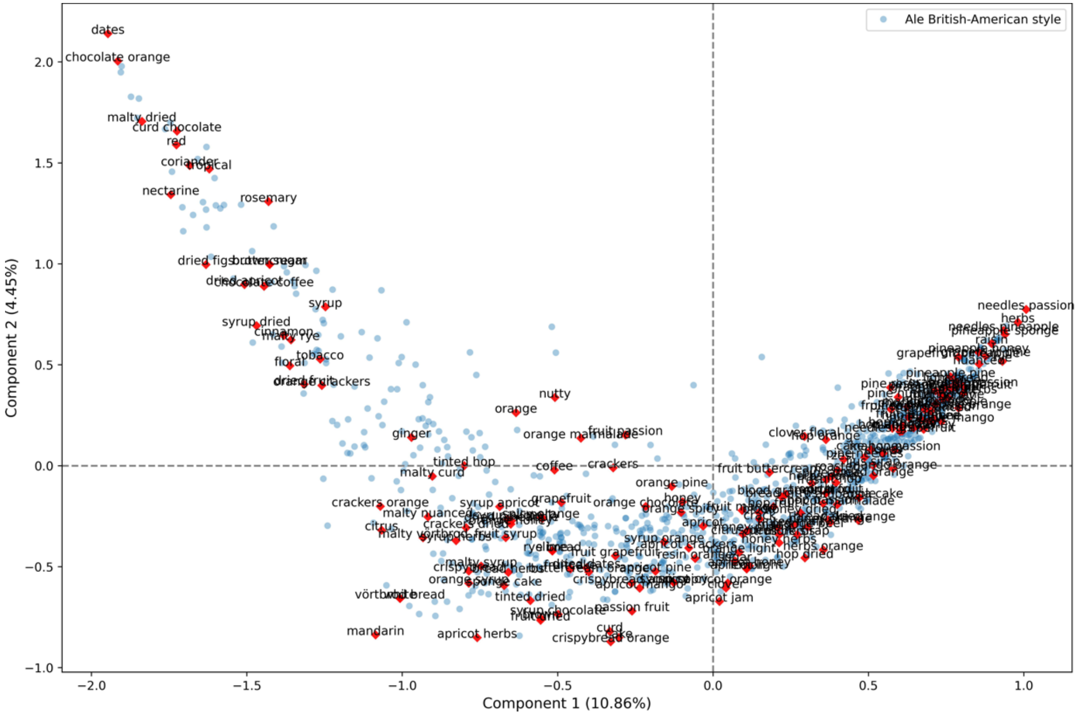
Common MVDA in wine sciences

Dimension reduction

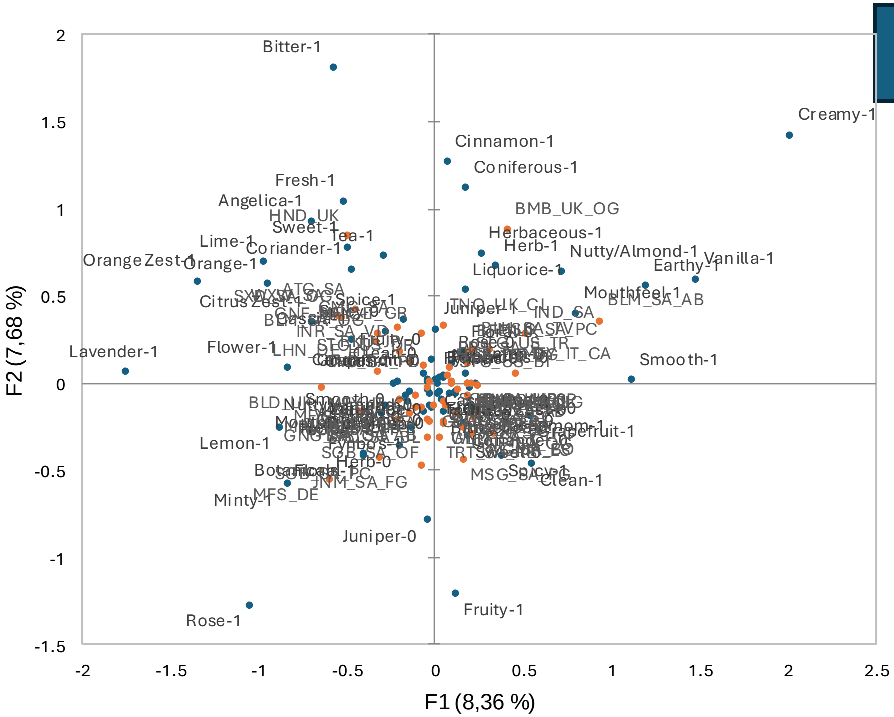
Some come with a common visualization, some do not. Ultimately, they are just **mathematical calculations**, applied according to the type of data (e.g., Categorical, continuous, discreet, nominal, sparse). You get to represent the results in a communicative way that elucidates insight.

1 Orthogonal decomposition

CA



MCA



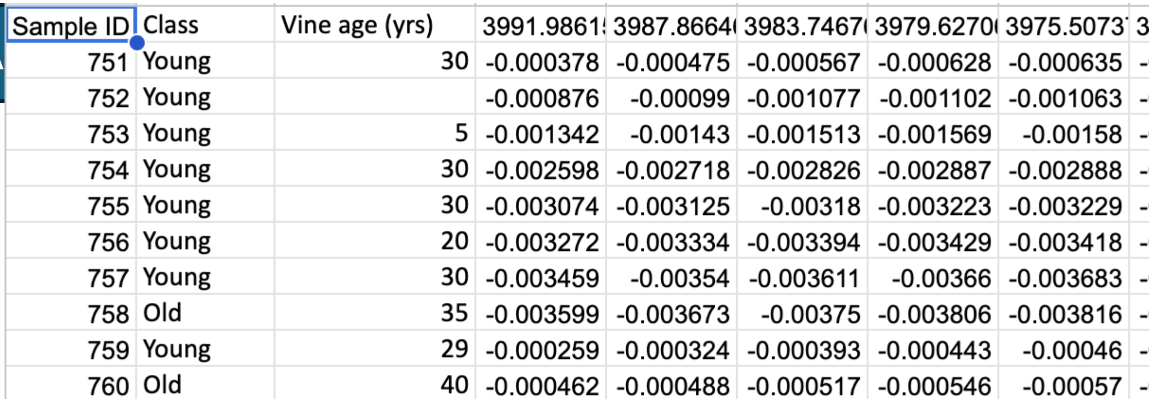
Common MVDA in wine sciences

Dimension reduction

Some come with a common visualization, some do not. Ultimately, they are just **mathematical calculations**, applied according to the type of data (e.g., Categorical, continuous, discreet, nominal, sparse). You get to represent the results in a communicative way that elucidates insight.

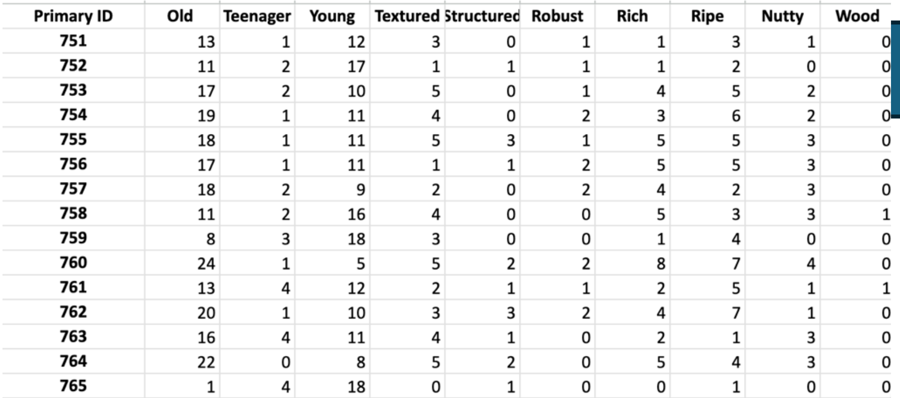
1 Orthogonal decomposition

PCA



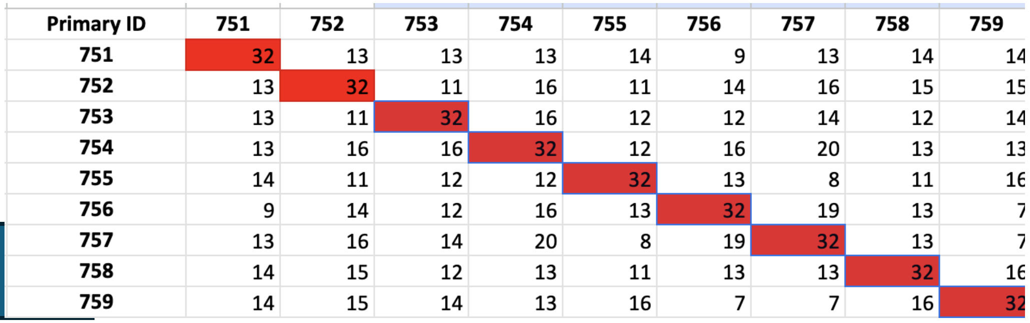
Sample ID	Class	Vine age (yrs)	3991.9861	3987.8664	3983.7467	3979.6270	3975.5073	3
751	Young	30	-0.000378	-0.000475	-0.000567	-0.000628	-0.000635	-
752	Young		-0.000876	-0.00099	-0.001077	-0.001102	-0.001063	-
753	Young	5	-0.001342	-0.00143	-0.001513	-0.001569	-0.00158	-
754	Young	30	-0.002598	-0.002718	-0.002826	-0.002887	-0.002888	-
755	Young	30	-0.003074	-0.003125	-0.00318	-0.003223	-0.003229	-
756	Young	20	-0.003272	-0.003334	-0.003394	-0.003429	-0.003418	-
757	Young	30	-0.003459	-0.00354	-0.003611	-0.00366	-0.003683	-
758	Old	35	-0.003599	-0.003673	-0.00375	-0.003806	-0.003816	-
759	Young	29	-0.000259	-0.000324	-0.000393	-0.000443	-0.00046	-
760	Old	40	-0.000462	-0.000488	-0.000517	-0.000546	-0.00057	-

CA



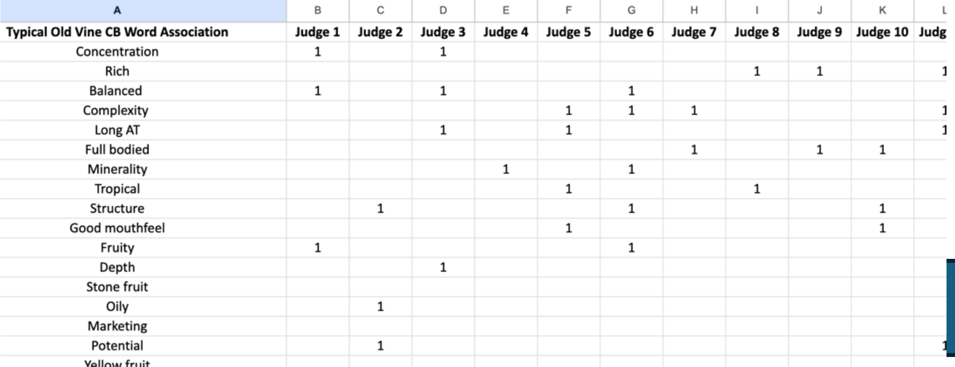
Primary ID	Old	Teenager	Young	Textured	Structured	Robust	Rich	Ripe	Nutty	Wood
751	13	1	12	3	0	1	1	3	1	0
752	11	2	17	1	1	1	1	2	0	0
753	17	2	10	5	0	1	4	5	2	0
754	19	1	11	4	0	2	3	6	2	0
755	18	1	11	5	3	1	5	5	3	0
756	17	1	11	1	1	2	5	5	3	0
757	18	2	9	2	0	2	4	2	3	0
758	11	2	16	4	0	0	5	3	3	1
759	8	3	18	3	0	0	1	4	0	0
760	24	1	5	5	2	2	8	7	4	0
761	13	4	12	2	1	1	2	5	1	1
762	20	1	10	3	3	2	4	7	1	0
763	16	4	11	4	1	0	2	1	3	0
764	22	0	8	5	2	0	5	4	3	0
765	1	4	18	0	1	0	0	1	0	0

MDS



Primary ID	751	752	753	754	755	756	757	758	759
751	32	13	13	13	14	9	13	14	14
752	13	32	11	16	11	14	16	15	15
753	13	11	32	16	12	12	14	12	14
754	13	16	16	32	12	16	20	13	13
755	14	11	12	12	32	13	8	11	16
756	9	14	12	16	13	32	19	13	7
757	13	16	14	20	8	19	32	13	7
758	14	15	12	13	11	13	13	32	16
759	14	15	14	13	16	7	7	16	32

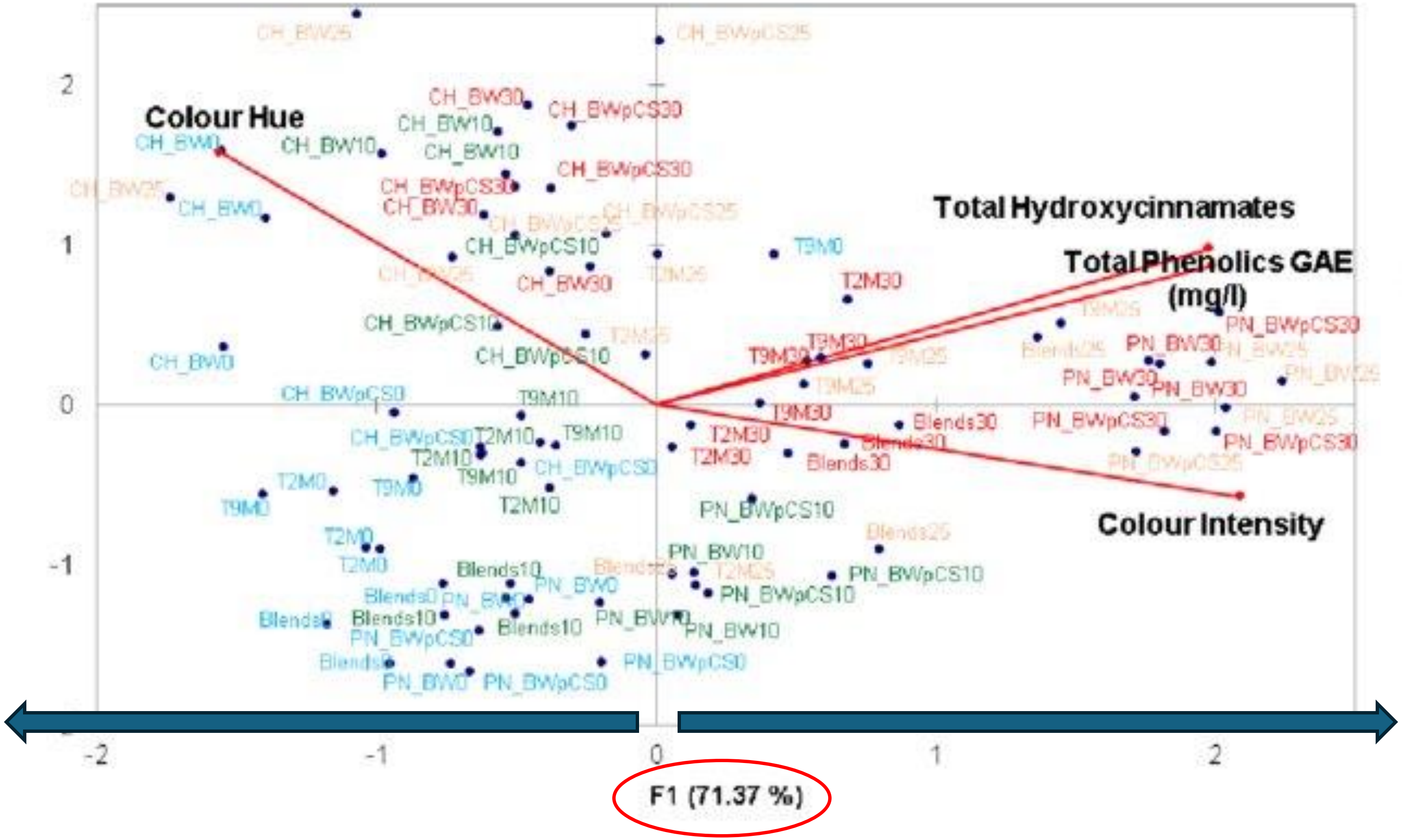
MCA



A	B	C	D	E	F	G	H	I	J	K	L
Typical Old Vine CB Word Association	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8	Judge 9	Judge 10	Judg
Concentration	1		1								
Rich								1	1		1
Balanced	1		1			1					
Complexity					1	1	1				1
Long AT			1		1						1
Full bodied							1		1	1	
Minerality				1		1					
Tropical					1			1			
Structure		1				1				1	
Good mouthfeel					1					1	
Fruity	1					1					
Depth			1								
Stone fruit											
Oily		1									
Marketing											
Potential		1									1
Yellow fruit											

Common MVDA in wine sciences

Reading biplots/cartesian covariance representations



Varies negatively to F1 = **lower**

Varies positively to F1 = **higher**

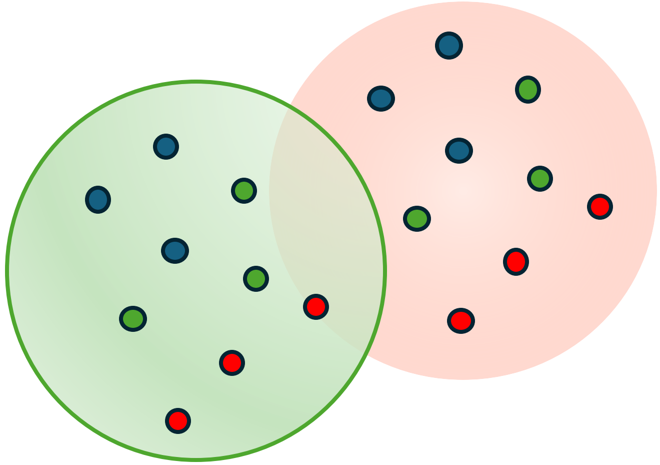
Data wrangling & Caveats in MVDA

- Aggregation method: sum, counts, etc.
- Treating NaNs – empty vs zeros (consequences?)
- Grouping

Wrangling libraries and techniques:
R - tidyverse
Python – pandas, numpy
SQL - count

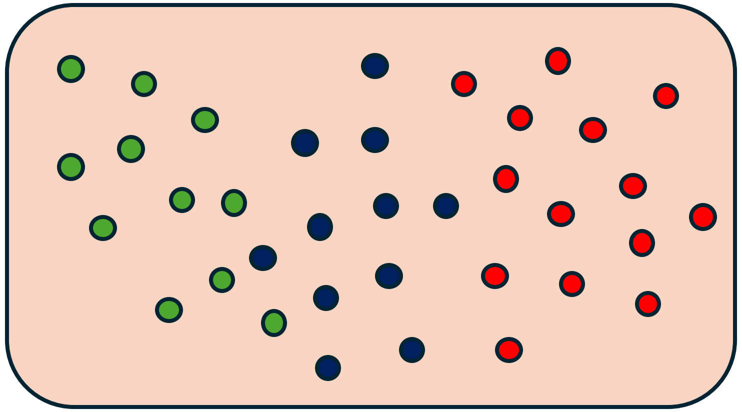
Grouping observations

The variance across observational groups of **samples** is greater than between each sample



Grouping variables

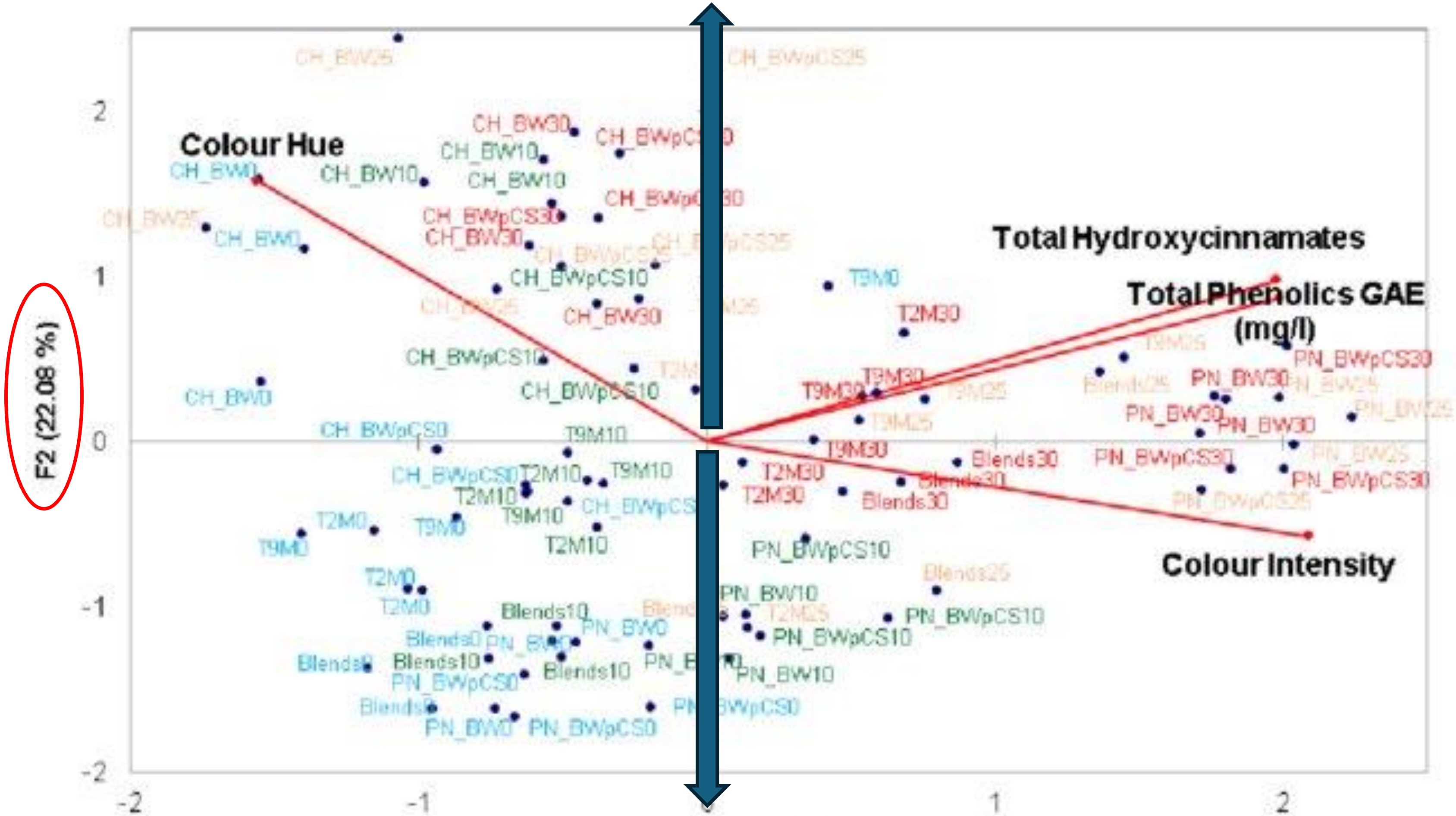
The variance across observational groups of **measurements** is greater than between each measurements



Common MVDA in wine sciences

Reading biplots/cartesian covariance representations

Varies positively to F2 = **higher**



Varies negatively to F2 = **lower**

Common MVDA in wine sciences

Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Common MVDA in wine sciences

ECO2WINE

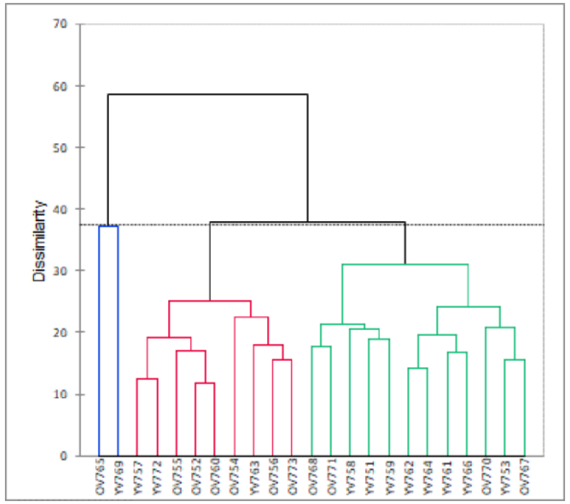
Cluster analysis
Often calculated on results from dimension reduction (recommended!).
e.g., Hierarchical clustering (HCA),

2

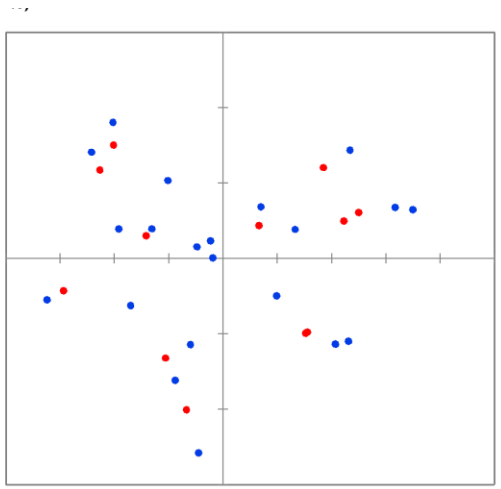
Cluster analysis

Number of clusters: 9

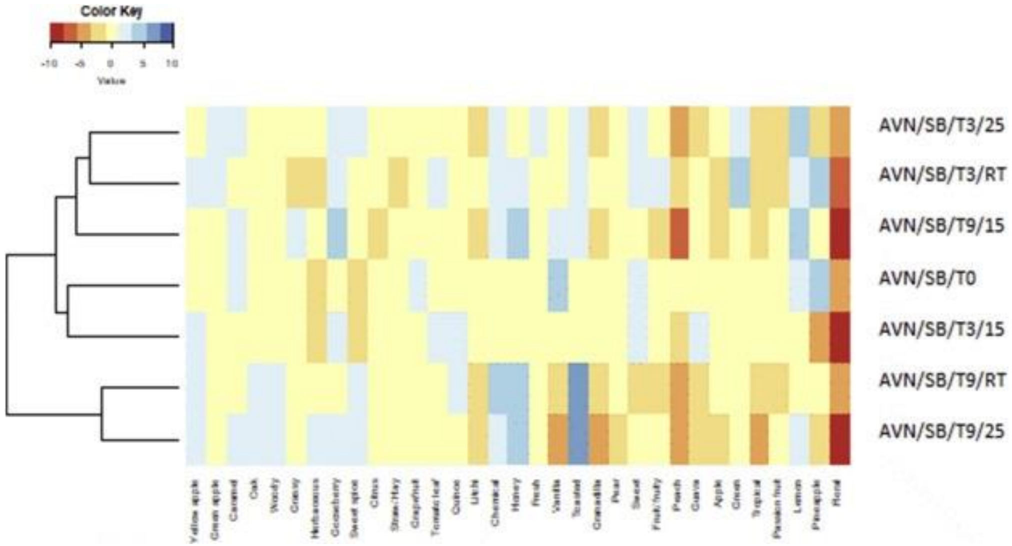
Cluster: 1 755
Cluster: 1 756
Cluster: 1 761
Cluster: 1 762
Cluster: 1 764
Cluster: 1 766
Cluster: 1 767
Cluster: 1 770
Cluster: 2 752
Cluster: 2 754
Cluster: 2 757
Cluster: 2 758
Cluster: 2 759
Cluster: 2 760
Cluster: 3 763
Cluster: 3 771
Cluster: 3 773
Cluster: 4 765
Cluster: 5 769
Cluster: 6 753
Cluster: 7 768
Cluster: 8 772
Cluster: 9 751



Dendrogram



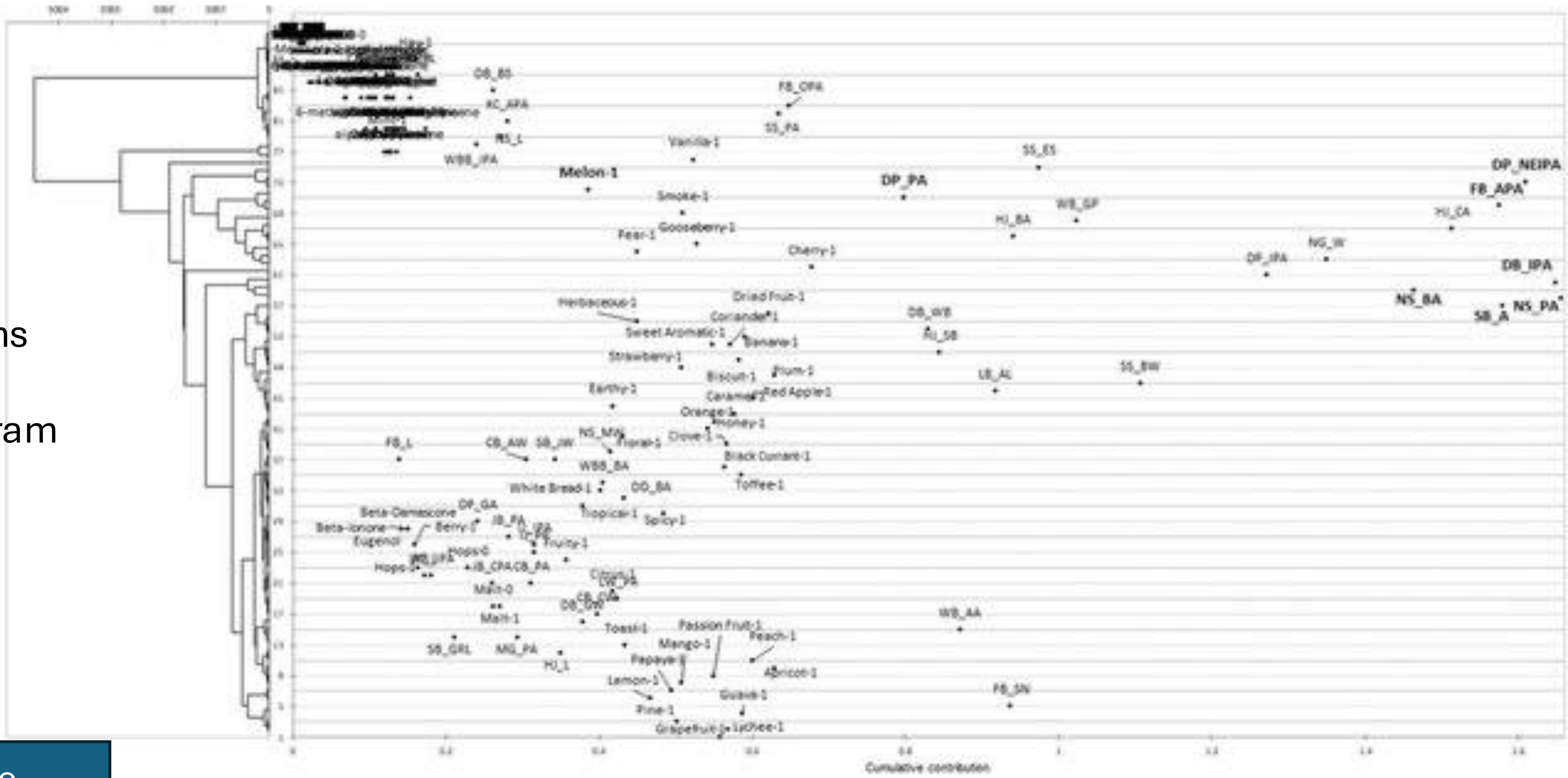
scatterplot



heatmap

Common MVDA in wine sciences

PCA correlations
+
Cluster dendrogram



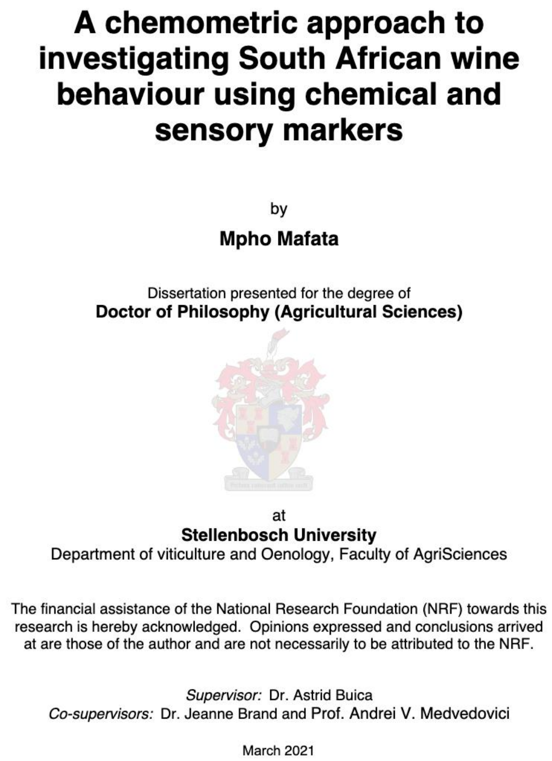
Just a scatterplot, so . . .

Thank you!

https://github.com/mpho-mafata

https://www.linkedin.com/in/mafatampho

https://orcid.org/0000-0002-6468-7193



A chemometric approach to investigating South African wine behaviour using chemical and sensory markers

by

Mpho Mafata

Dissertation presented for the degree of
Doctor of Philosophy (Agricultural Sciences)

at
Stellenbosch University
Department of viticulture and Oenology, Faculty of AgriSciences

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

Supervisor: Dr. Astrid Buica
Co-supervisors: Dr. Jeanne Brand and Prof. Andrei V. Medvedovici

March 2021

https://doi.org/10.13140/RG.2.2.17899.00804