Unsupervised Learning Capstone

Clustering Palm Tree Species

Abhishek Verma

Table of Content

- Overview
- Objective
- Data Source
- Data Exploration: Palm Subfamily
- Dimensionality Reduction Analysis
- Unsupervised Modelling
- Modelling Evaluation
- Limitation
- Conclusion

Presentation Duration : 20 min

Overview

- Palm tree are among the most cultivated plant around the globe. They are economically important in providing byproducts in food, wood, construction, medicine etc.

- Different species of palm trees are grown globally with varying physical traits like stem size, leaves, fruit size etc. that determine their commercial value in the market.

- Climate, region and ecosystem plays an important role in the growth and survival of these species.

Table of Content

- Overview
- Objective
- Data Source
- Data Exploration: Palm Subfamily
- Dimensionality Reduction Analysis
- Unsupervised Modelling
- Modelling Evaluation
- Limitation
- Conclusion

Presentation Duration : 20 min

Objective: Unsupervised Modelling

- To form clusters unsupervised that share common physical traits of palm tree species.

- To test the hypothesis that the clustering based on plants traits or physical properties is in agreement with the taxonomic group ‘subfamily’.

Table of Content

- Overview
- Objective
- Data Source
- Data Exploration: Palm Subfamily
- Dimensionality Reduction Analysis
- Unsupervised Modelling
- Modelling Evaluation
- Limitation
- Conclusion

Presentation Duration : 20 min

Data Source

- The dataset used for this exercise is based on the derived measurement of over 2500 species of palm tree.

- Contains the global species compilation of functional traits of palm tree (Arecaceae) which predominantly grows in tropical and subtropical ecosystem

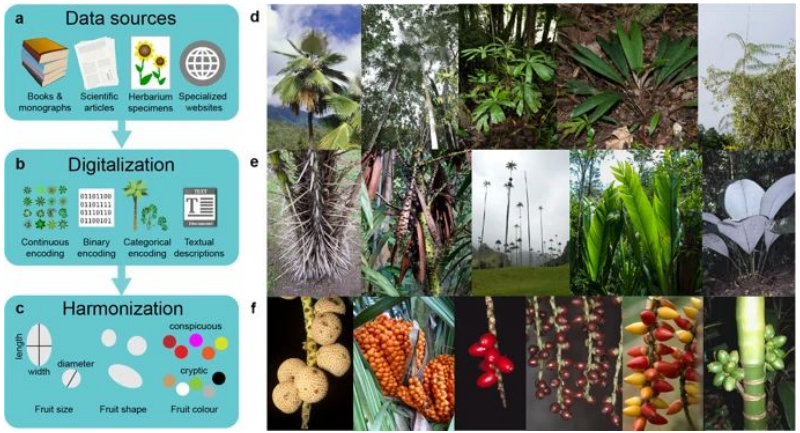
https://www.nature.com/articles/s41597-019-0189-0#Tab1

Data Descriptor | Open Access | Published: 24 September 2019

PalmTraits 1.0, a species-level functional trait database of palms worldwide

W. Daniel Kissling ✉, Henrik Balslev, William J. Baker, John Dransfield, Bastian Göldel, Jun Ying Lim, Renske E. Onstein & Jens-Christian Svenning

Scientific Data **6**, Article number: 178 (2019) | Cite this article



Data Source: Key Information

Taxonomy

Genus
Species
Palm Tribe
Palm Subfamily (count: 5)

Growth Form and Habit

Climbing
Acaulescent
Erect
StemSolitary

Stem Size

MaxStemHeight_m
MaxStemDia_cm
UnderstoreyCanopy

Armature

StemArmed
LeavesArmed

Fruit Properties

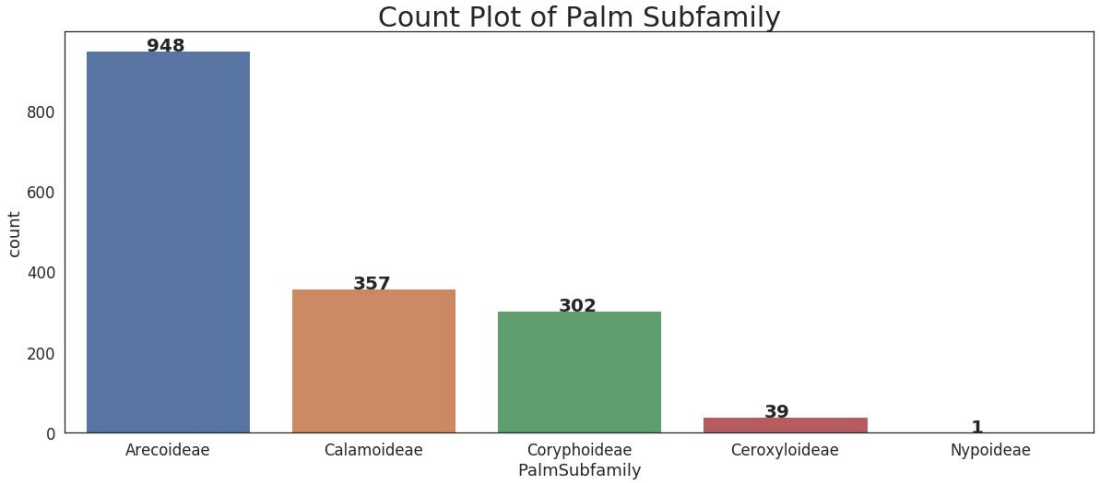
AverageFruitLength_cm
AverageFruitWidth_cm
FruitSizeCategorical

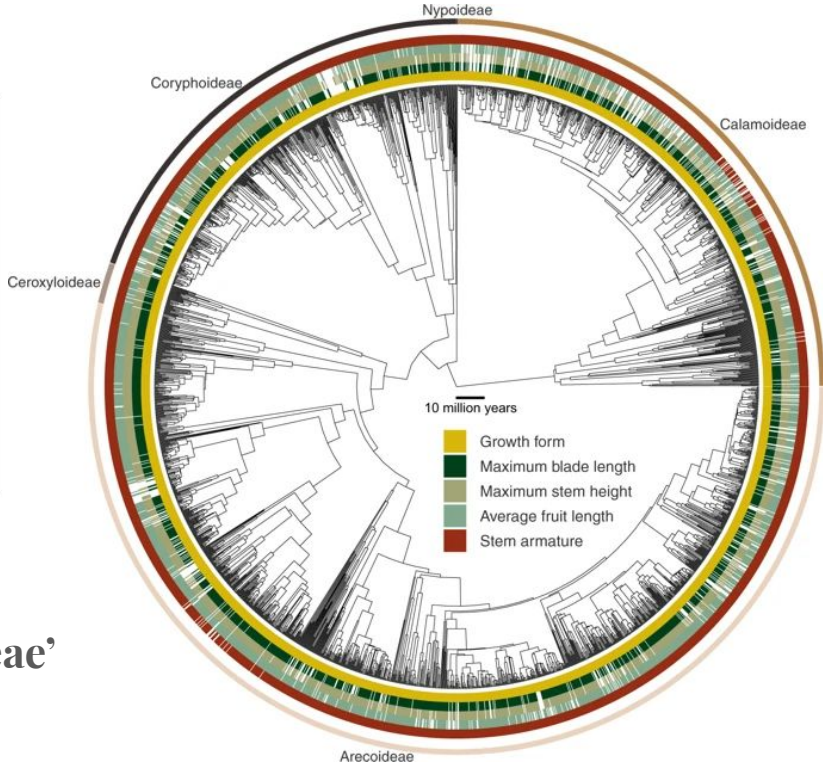
Table of Content

- Overview
- Objective
- Data Source
- Data Exploration: Palm Subfamily
- Dimensionality Reduction Analysis
- Unsupervised Modelling
- Modelling Evaluation
- Limitation
- Conclusion

Presentation Duration : 20 min

Palm Subfamily: Distribution

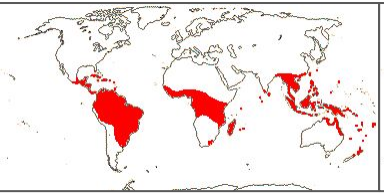
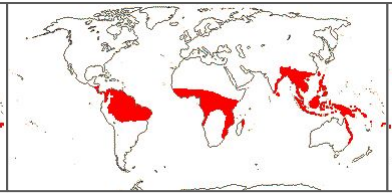
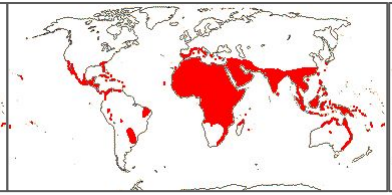
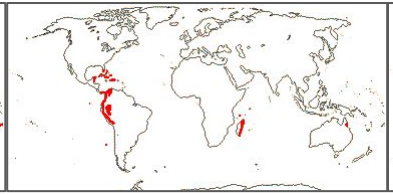
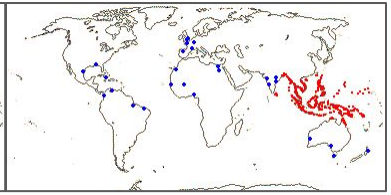




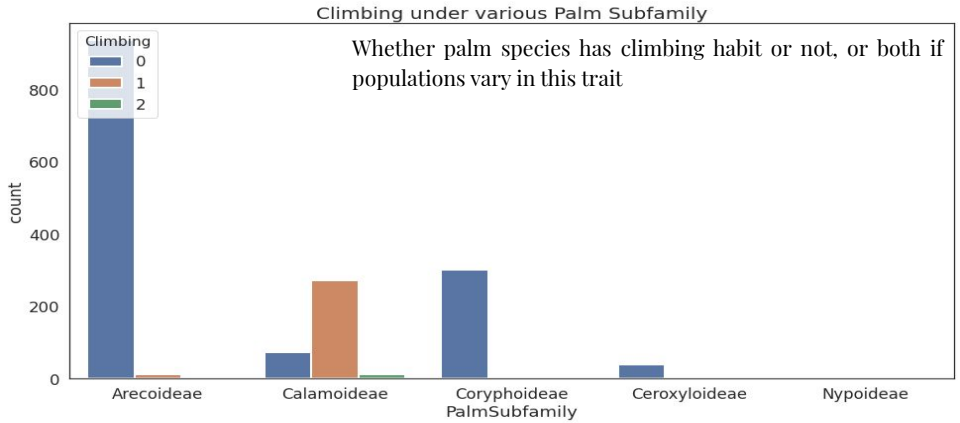
- Maximum number of species comes from **'Arecoideae'** subfamily

https://www.nature.com/articles/s41597-019-0189-0#Tab1

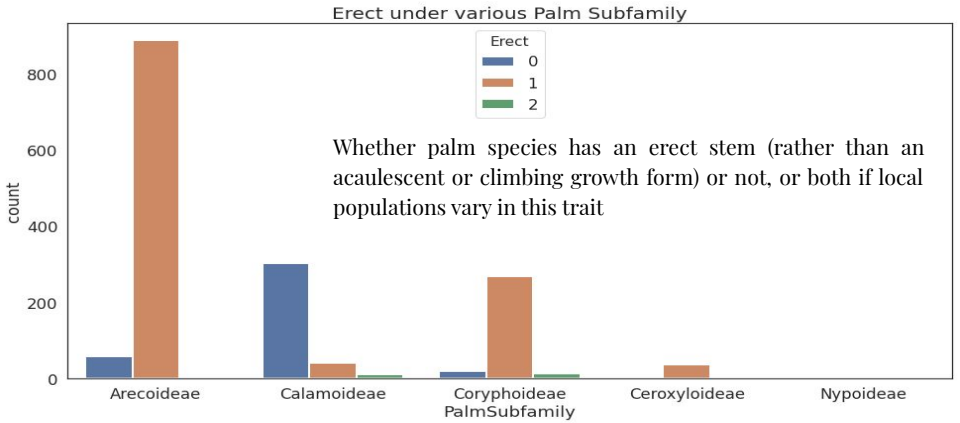
Palm Subfamily: Example

Arecoideae	Calamoideae	Coryphoideae	Ceroxyloideae	Nypoideae
				
Cocos nucifera	*Raphia farinifera*	*Schippia*	*Ceroxylon*	*Nypa Fruticans*
				

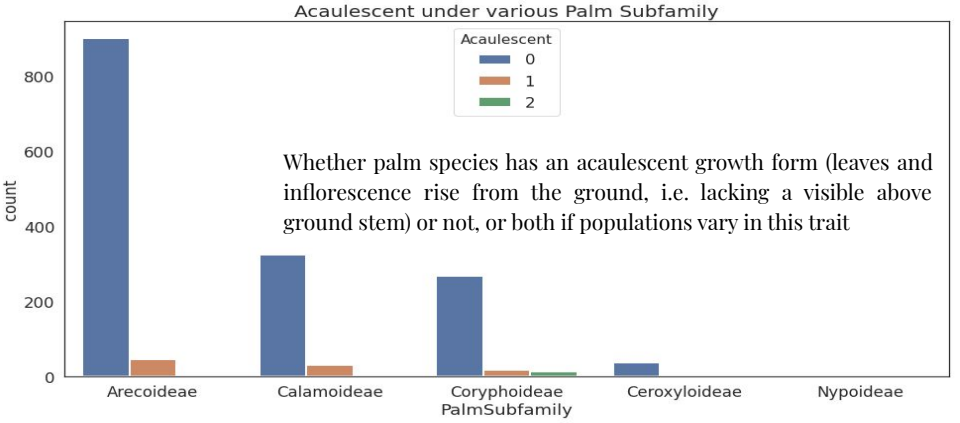
Palm Subfamily: Growth Form and Habit



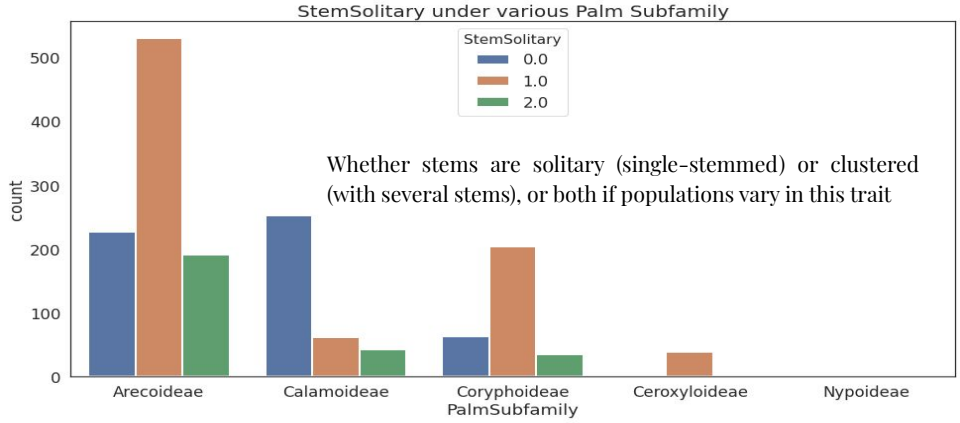
Whether palm species has climbing habit or not, or both if populations vary in this trait



Whether palm species has an erect stem (rather than an acaulescent or climbing growth form) or not, or both if local populations vary in this trait

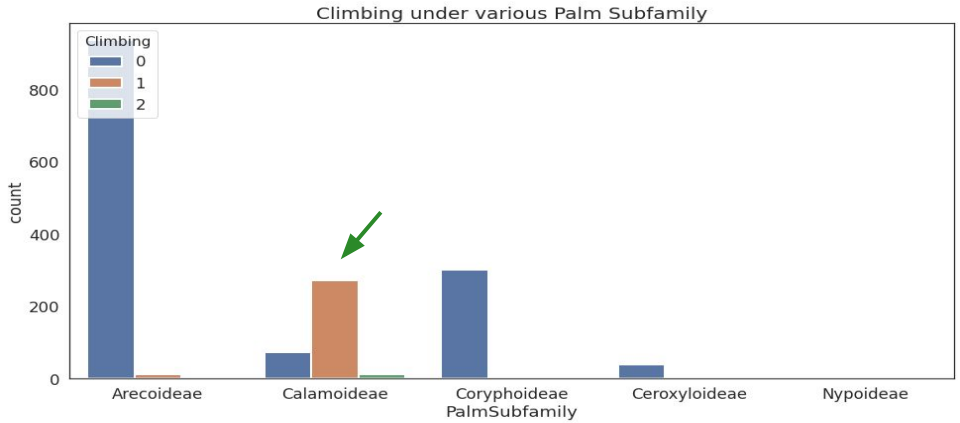


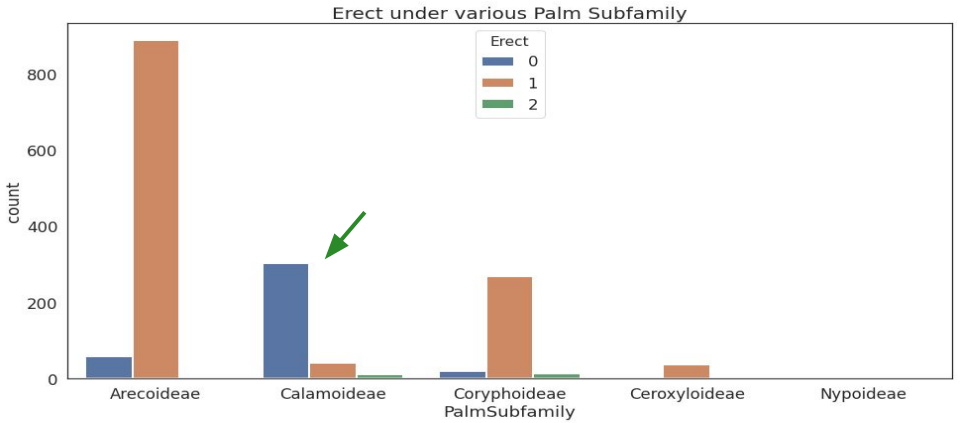
Whether palm species has an acaulescent growth form (leaves and inflorescence rise from the ground, i.e. lacking a visible above ground stem) or not, or both if populations vary in this trait

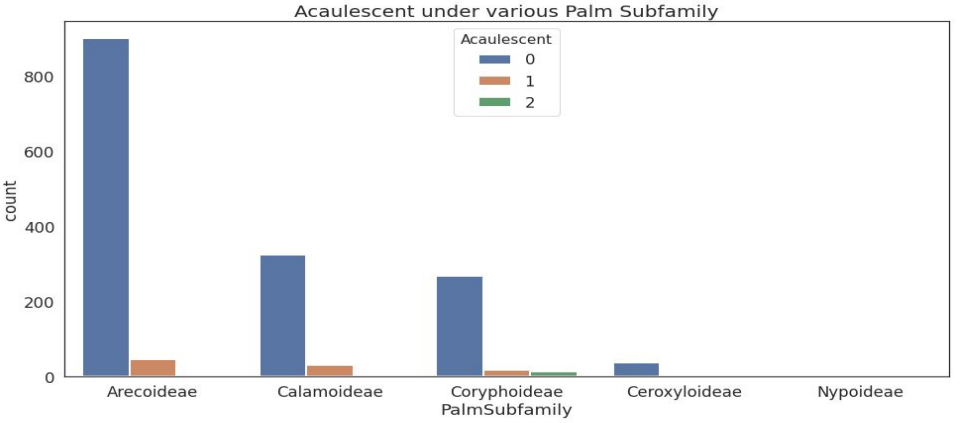


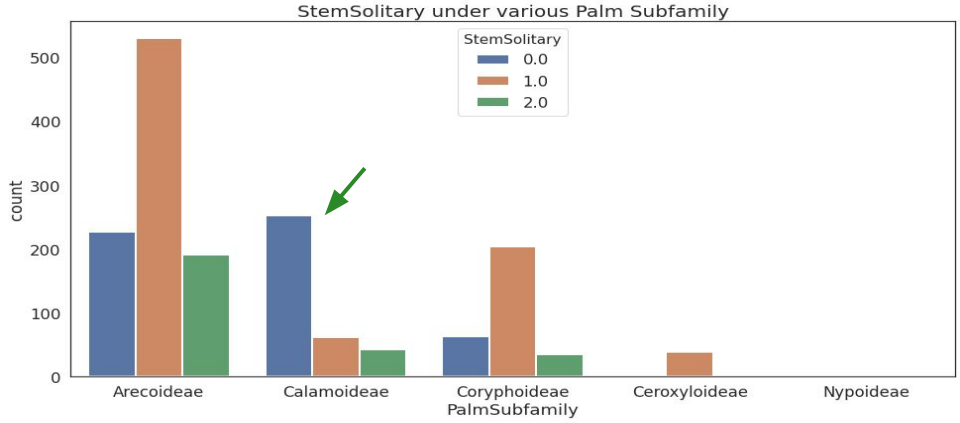
Whether stems are solitary (single-stemmed) or clustered (with several stems), or both if populations vary in this trait

Palm Subfamily: Growth Form and Habit

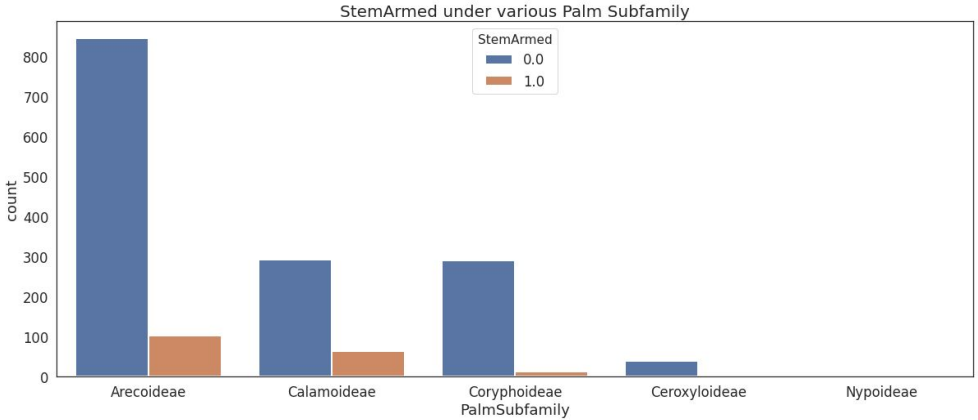




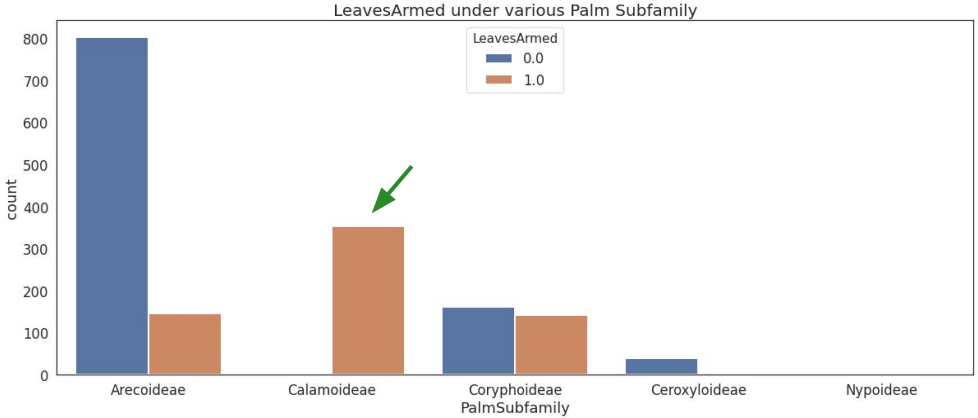




Palm Subfamily: Armature

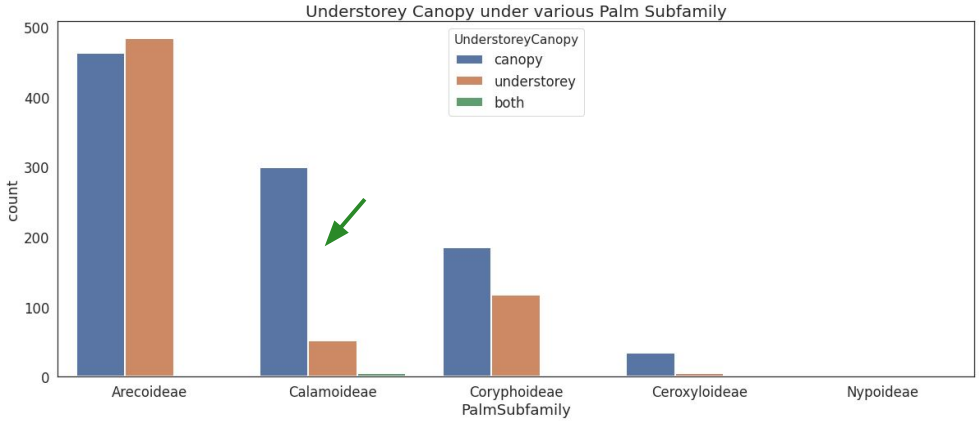


StemArmed: Whether bearing some form of spines at the stem or not, or both if populations vary in this trait



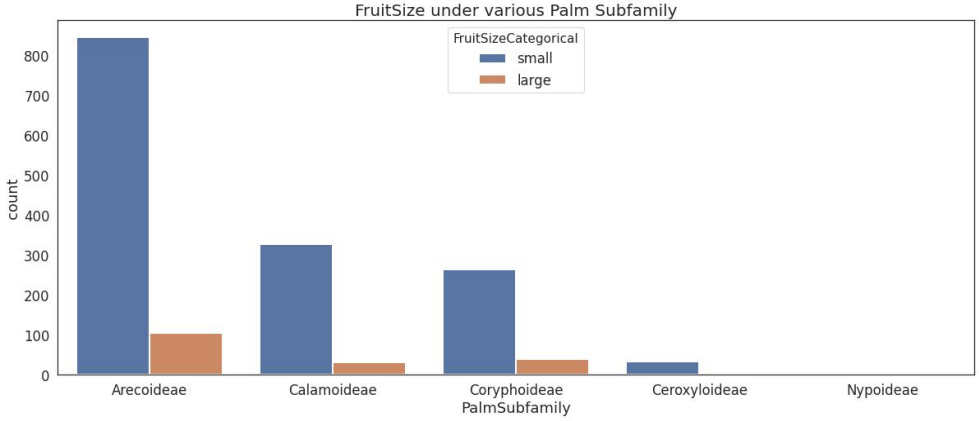
LeavesArmed: Whether bearing some form of spines on the leaves or not, or both if populations vary in this trait

Categorical Variable



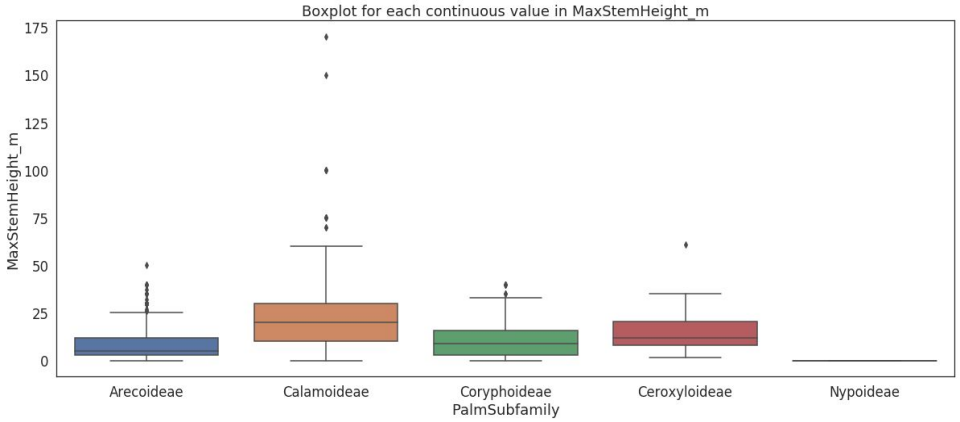
Understorey palms : short-stemmed palms with a maximum stem height ≤ 5 m or an acaulescent growth form,

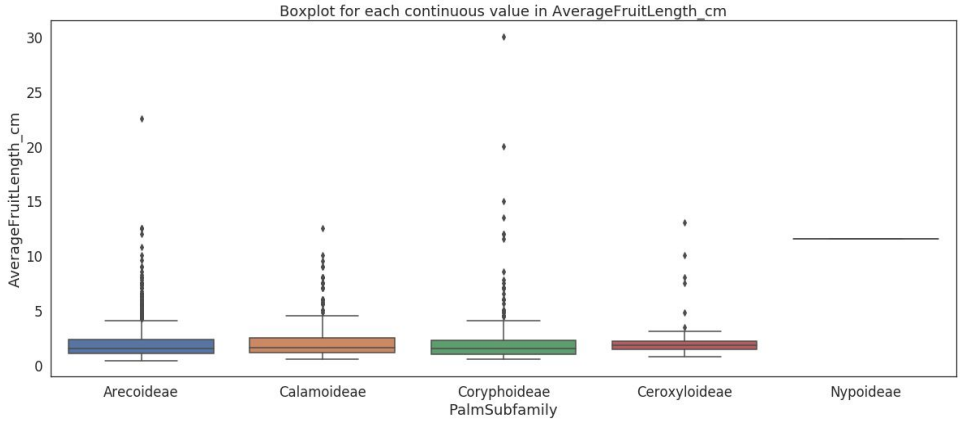
Canopy palms: maximum stem height > 5

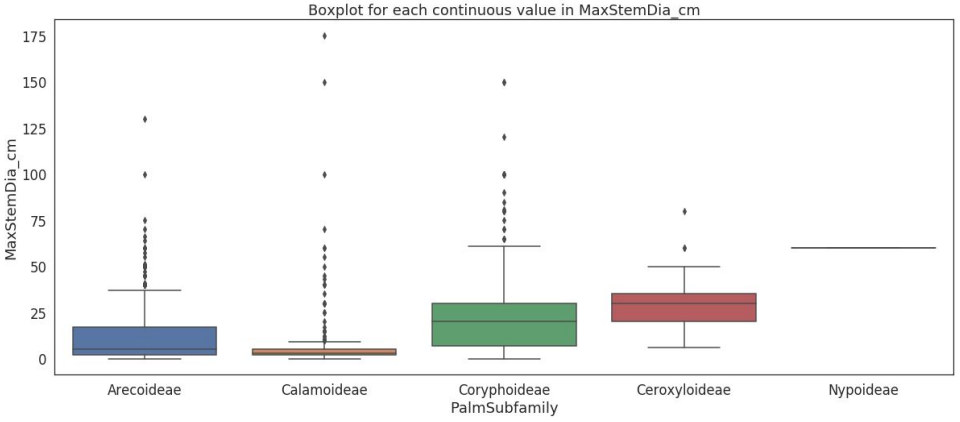


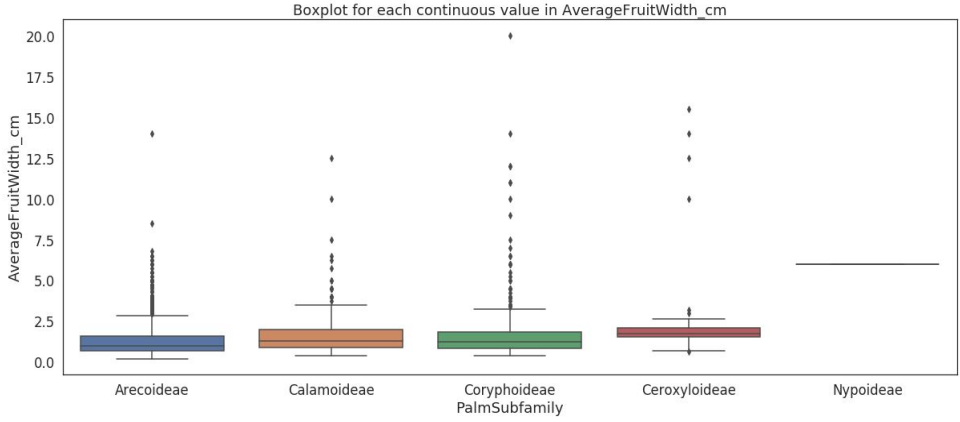
Small-fruited palms : fruits < 4 cm in length
Large-fruited palms : fruits ≥ 4 cm in length

Bivariate Analysis: Stems and Fruit Properties

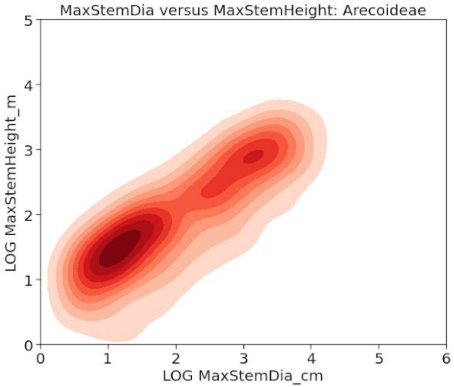


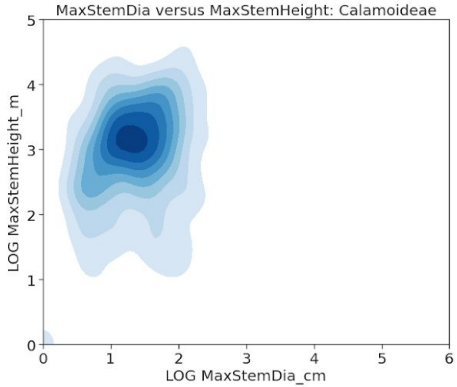


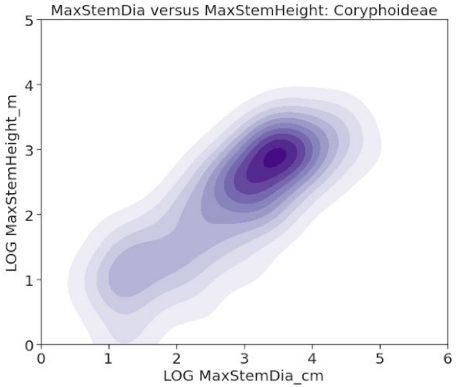


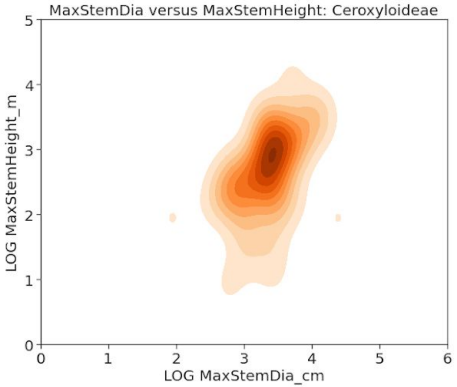


Bivariate Analysis: Stems and Fruit Properties

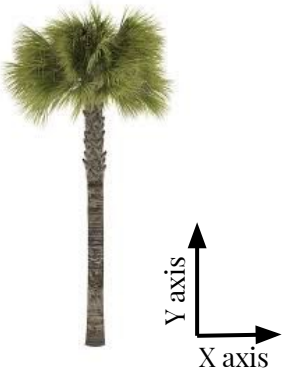




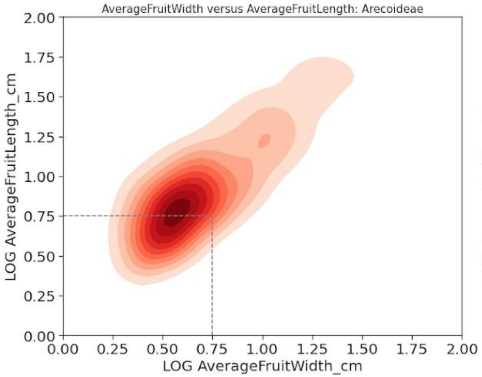


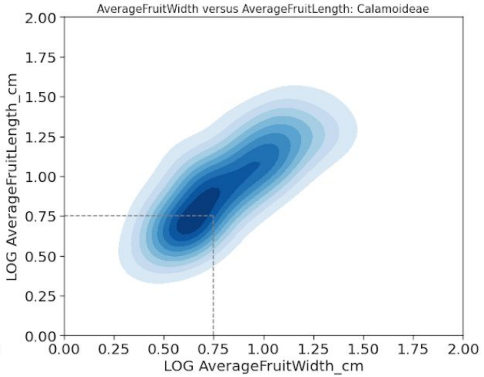


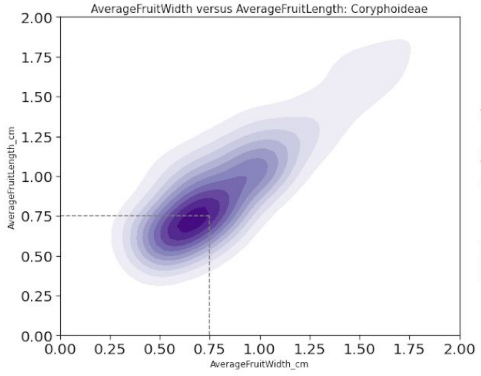
- Differences and similarity in distribution pattern of stem height vs stem thickness is show
- Eg: Calamoideae trees are overall taller and thinner versus Ceroxyloideae trees that are taller and broader

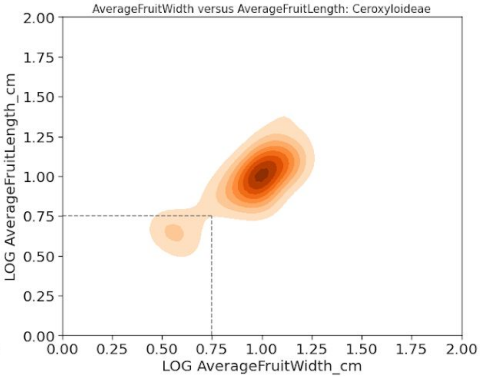


Bivariate Analysis: Stems and Fruit Properties









- Ceroxyloideae trees have bigger fruits sizes.

Bivariate: Stem versus fruit properties

- Average fruit length and fruit width are highly correlated.

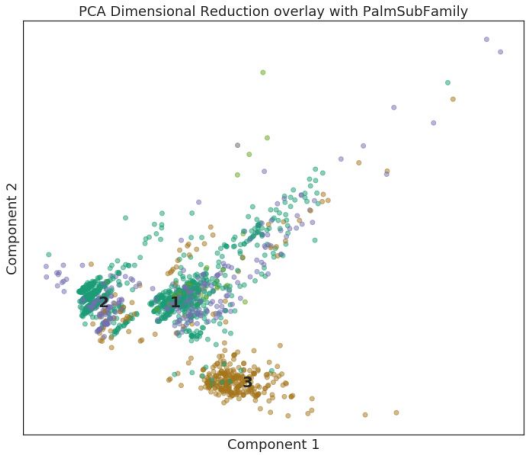
- Stem diameter to some degree is correlated with fruit sizes.

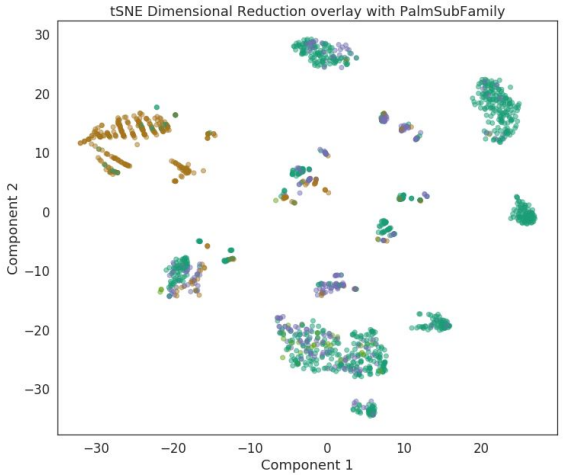
Table of Content

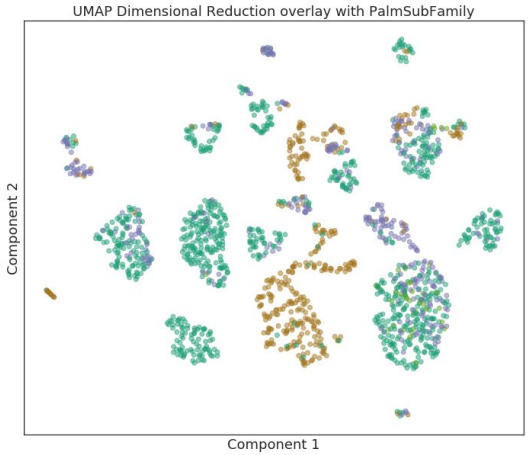
- Overview
- Objective
- Data Source
- Data Exploration: Palm Subfamily
- Dimensionality Reduction Analysis
- Unsupervised Modelling
- Modelling Evaluation
- Limitation
- Conclusion

Presentation Duration : 20 min

Dimensionality Reduction







- PCA, shows three linearly separated clusters that is reminiscent of the three major palm subfamily.
- T-SNE shows five large cluster that are linearly separable along with many small clusters with low data population. Similarly UMAP method shows five or more major cluster surrounded by scattered small clusters.

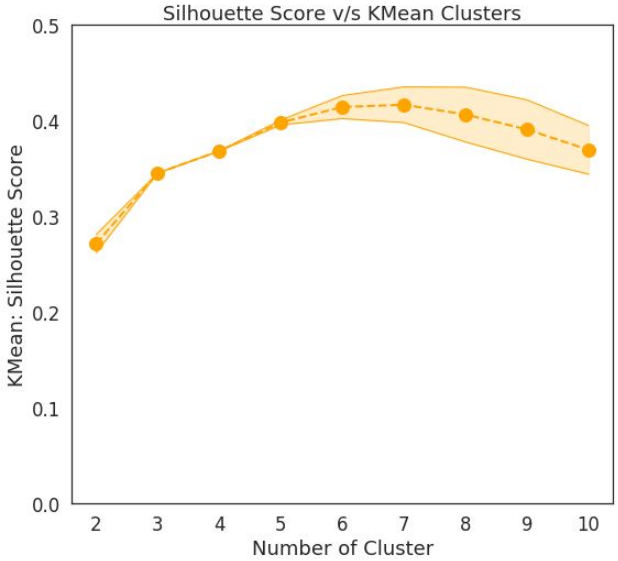
Table of Content

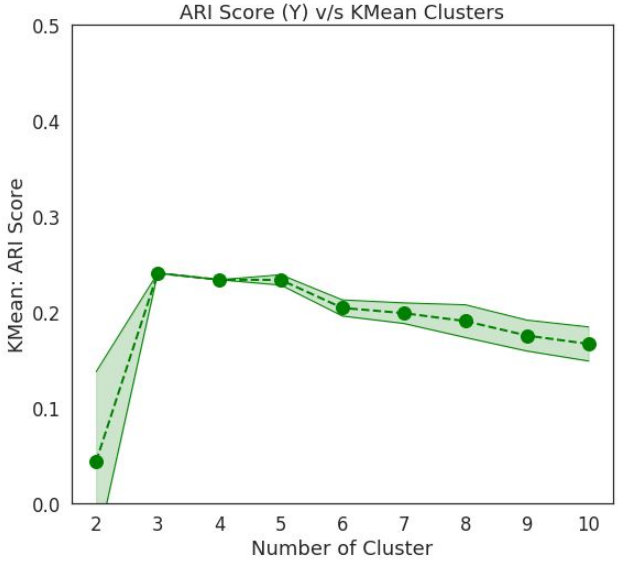
- Overview
- Objective
- Data Source
- Data Exploration: Palm Subfamily
- Dimensionality Reduction Analysis
- Unsupervised Modelling
- Modelling Evaluation
- Limitation
- Conclusion

Presentation Duration : 20 min

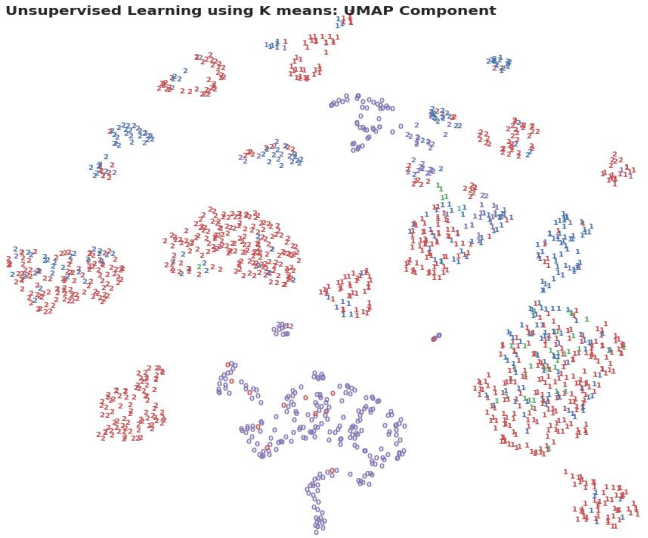
- **Silhouette Score**: Represent the cluster prediction in unsupervised fashion
- **ARI Score (Y)**: Classification of clusters based on taxonomy PalmSubfamily

K-Means Clustering





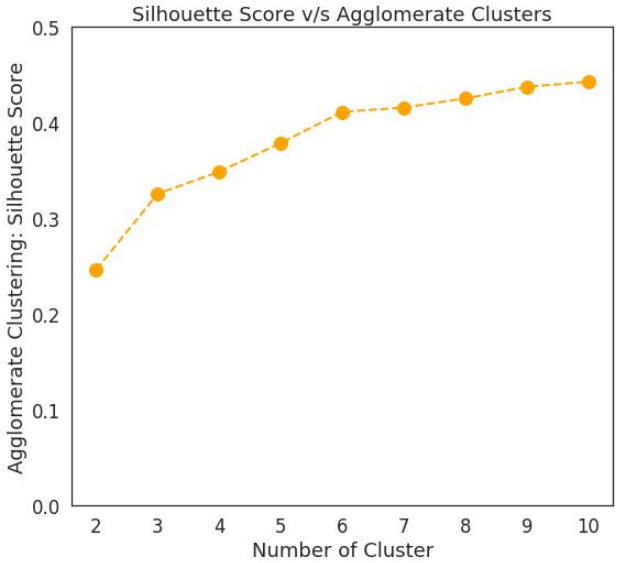
Iterated over 50 random states

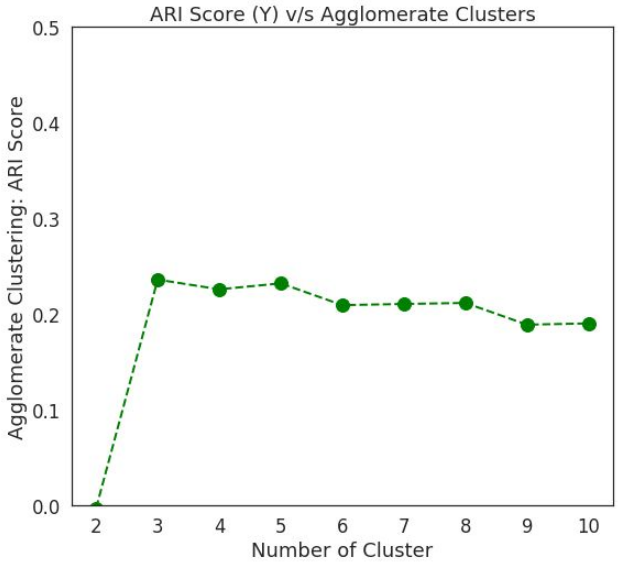


Color Code : Palm subfamily
Num Code : Modelling

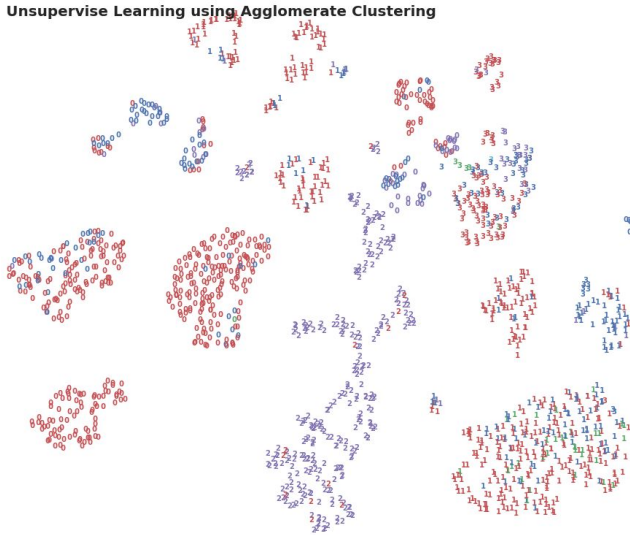
- Maximum Silhouette Score: 0.4 for 6-7 cluster
- ARI Score Y: 0.24 for n_cluster: 3

Hierarchical Clustering: Agglomerate





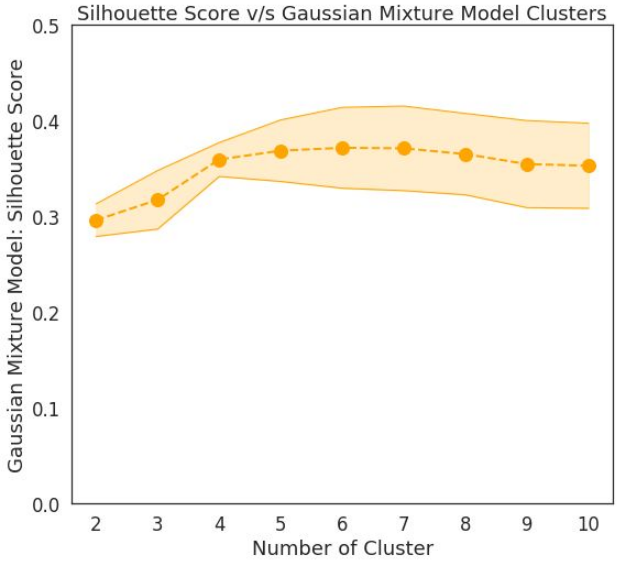
Iterated over 50 random states

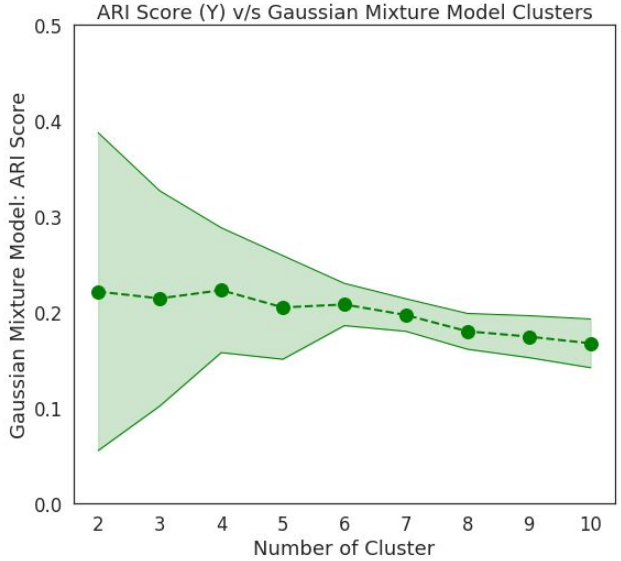


Color Code : Palm subfamily
Num Code : Modelling

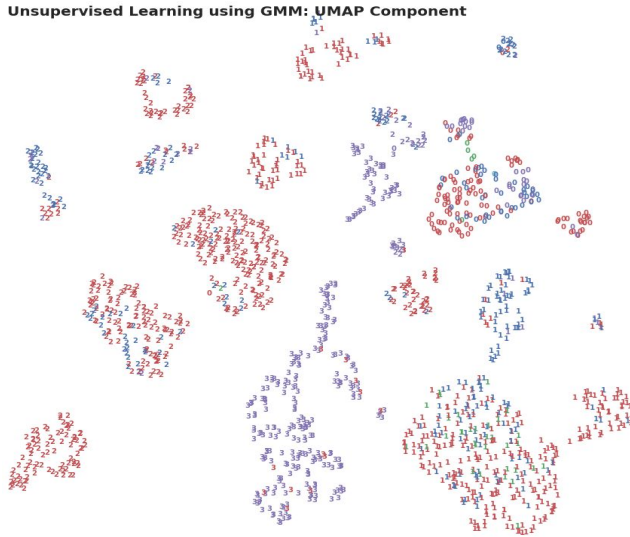
- Maximum Silhouette Score: 0.42 for 10 cluster (max limit of testing)
- ARI Score Y: 0.22 for n_cluster: 3

Gaussian Mixture Model Clustering





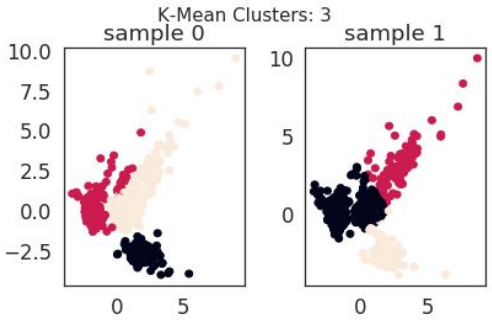
Iterated over 50 random states

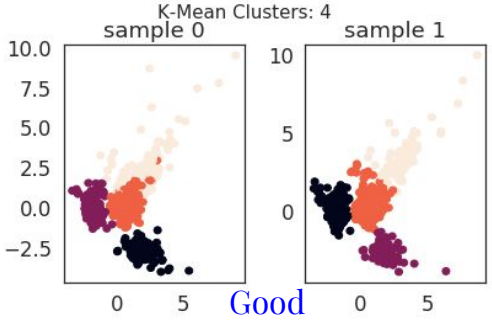


Color Code : Palm subfamily
Num Code : Modelling

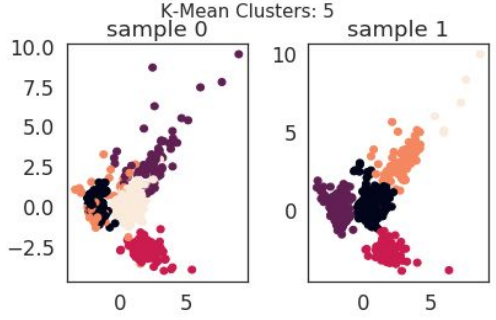
- Maximum Silhouette Score: 0.36 for n_cluster: 6
- ARI Score Y: 0.23 for n_cluster: 3

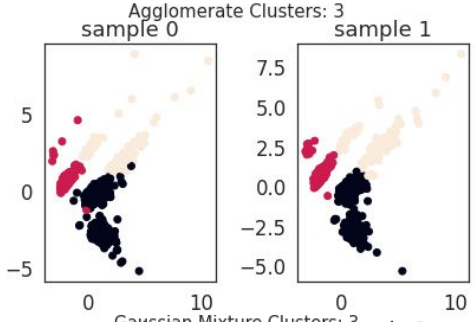
Consistency Test: Unspervised Models

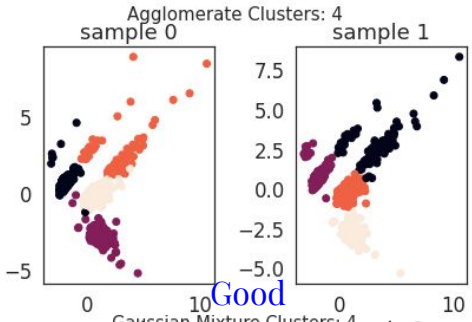




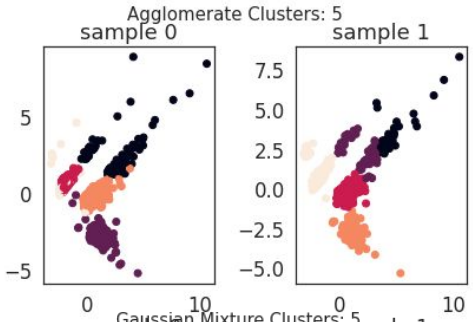
Good

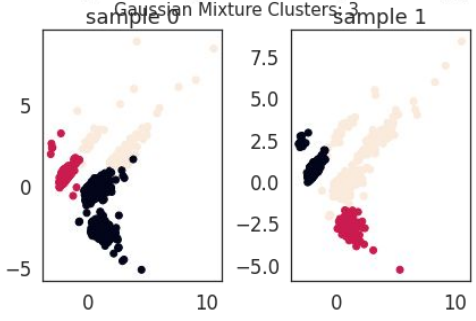


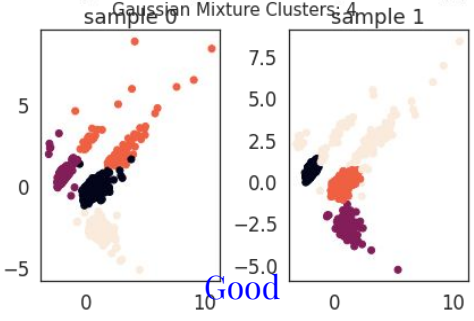




Good







Good

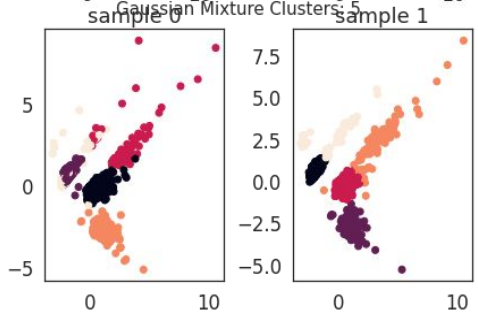
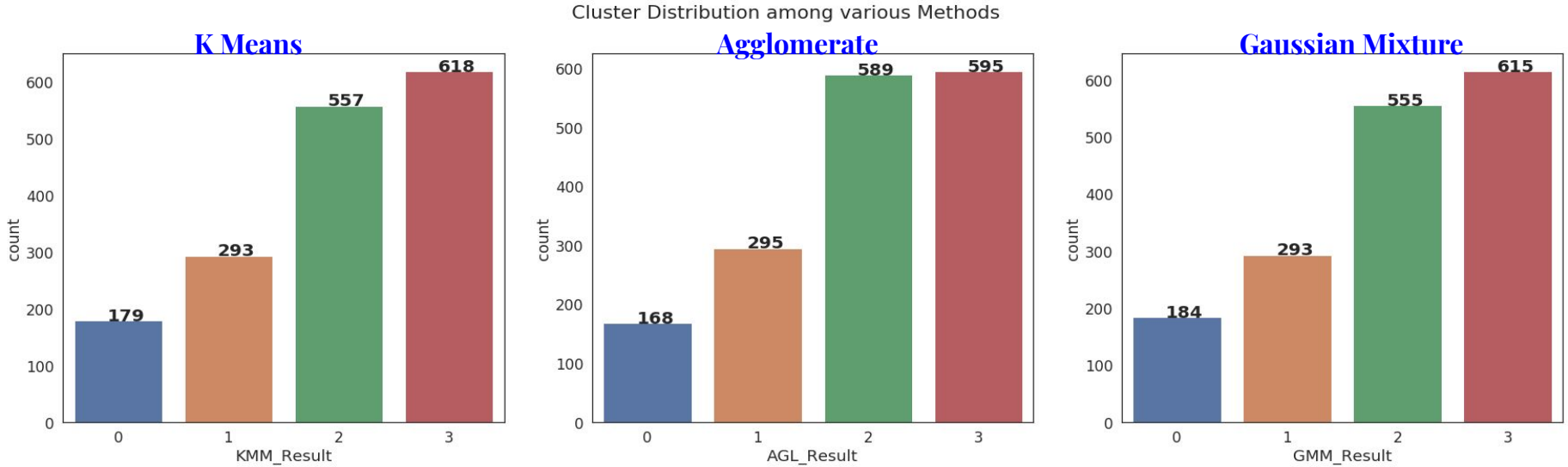


Table of Content

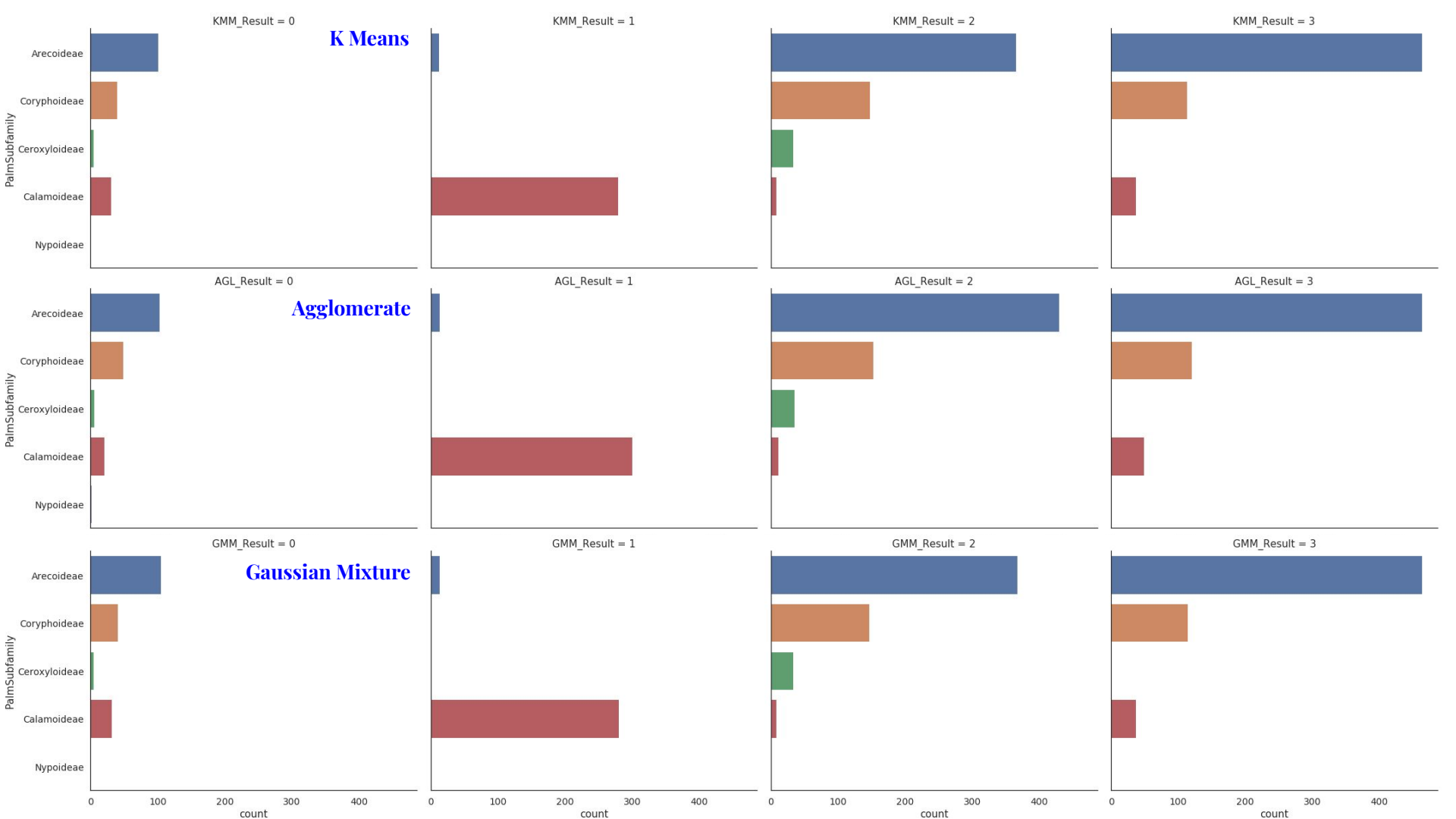
- Overview
- Objective
- Data Source
- Data Exploration: Palm Subfamily
- Dimensionality Reduction Analysis
- Unsupervised Modelling
- Modelling Evaluation
- Limitation
- Conclusion

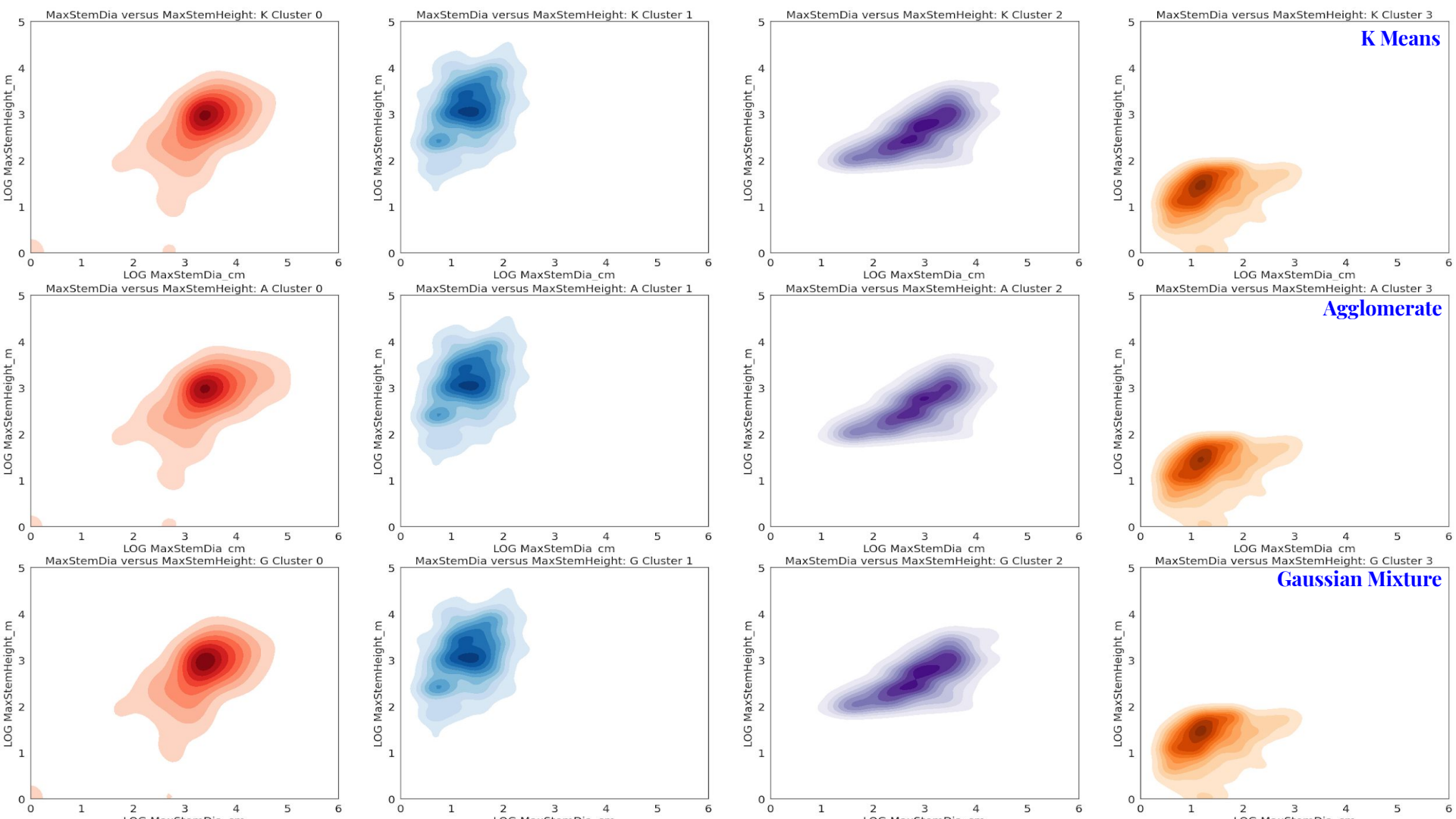
Presentation Duration : 20 min

Model Prediction: Number of Data points



- Almost consistent clustering among different unsupervised methods


KMM_Result = 0
KMM_Result = 1
KMM_Result = 2
KMM_Result = 3
K Means
AGL_Result = 0
AGL_Result = 1
AGL_Result = 2
AGL_Result = 3
Agglomerate
GMM_Result = 0
GMM_Result = 1
GMM_Result = 2
GMM_Result = 3
Gaussian Mixture
PalmSubfamily
Arecoideae
Coryphoideae
Ceroxyloideae
Calamoideae
Nypoideae
0
100
200
300
400
count


MaxStemDia versus MaxStemHeight: K Cluster 0
MaxStemDia versus MaxStemHeight: K Cluster 1
MaxStemDia versus MaxStemHeight: K Cluster 2
MaxStemDia versus MaxStemHeight: K Cluster 3
K Means
MaxStemDia versus MaxStemHeight: A Cluster 0
MaxStemDia versus MaxStemHeight: A Cluster 1
MaxStemDia versus MaxStemHeight: A Cluster 2
MaxStemDia versus MaxStemHeight: A Cluster 3
Agglomerate
MaxStemDia versus MaxStemHeight: G Cluster 0
MaxStemDia versus MaxStemHeight: G Cluster 1
MaxStemDia versus MaxStemHeight: G Cluster 2
MaxStemDia versus MaxStemHeight: G Cluster 3
Gaussian Mixture
LOG MaxStemHeight_m
LOG MaxStemDia_cm

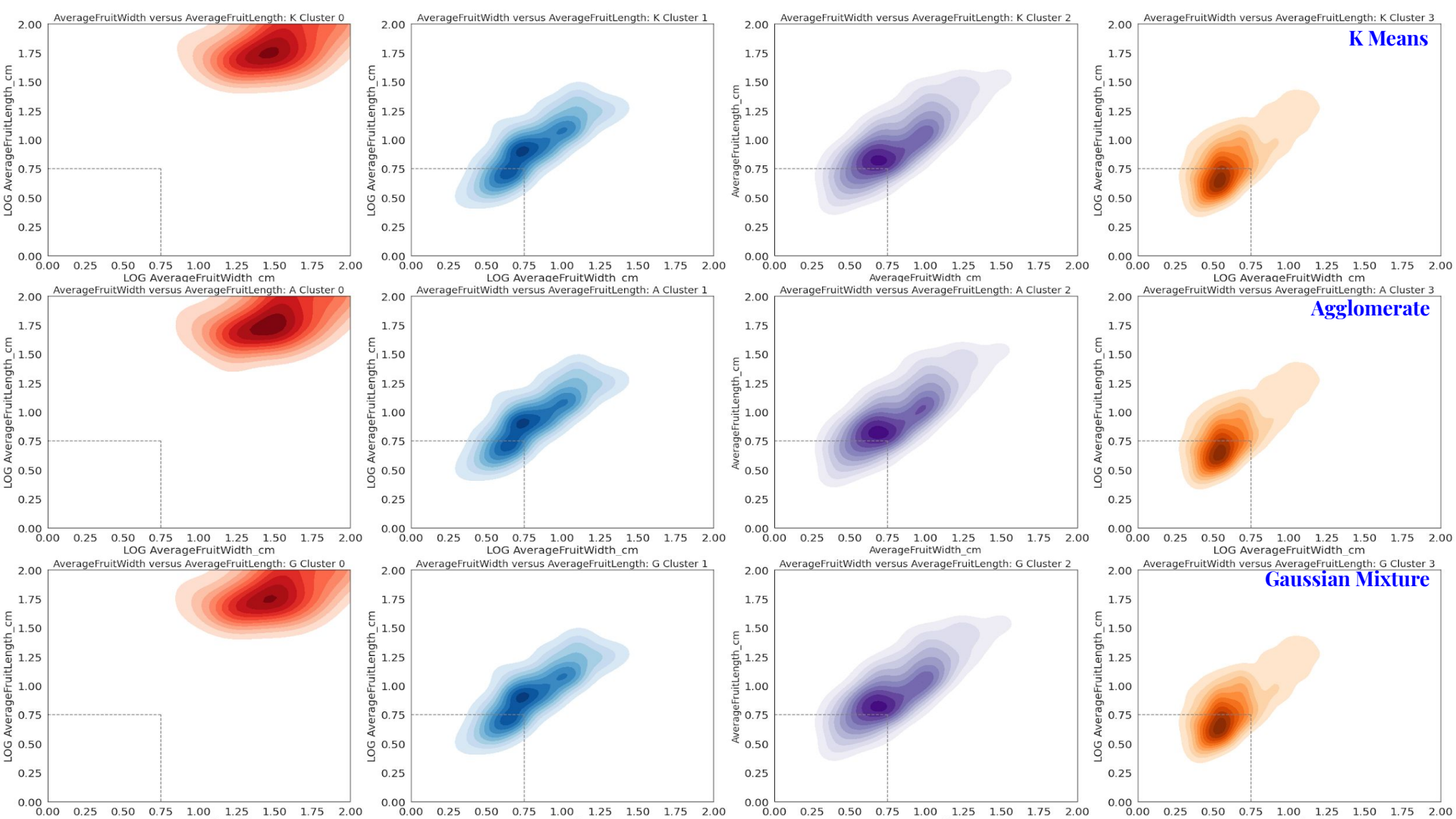

AverageFruitWidth versus AverageFruitLength: K Cluster 0
AverageFruitWidth versus AverageFruitLength: K Cluster 1
AverageFruitWidth versus AverageFruitLength: K Cluster 2
AverageFruitWidth versus AverageFruitLength: K Cluster 3
K Means
AverageFruitWidth versus AverageFruitLength: A Cluster 0
AverageFruitWidth versus AverageFruitLength: A Cluster 1
AverageFruitWidth versus AverageFruitLength: A Cluster 2
AverageFruitWidth versus AverageFruitLength: A Cluster 3
Agglomerate
AverageFruitWidth versus AverageFruitLength: G Cluster 0
AverageFruitWidth versus AverageFruitLength: G Cluster 1
AverageFruitWidth versus AverageFruitLength: G Cluster 2
AverageFruitWidth versus AverageFruitLength: G Cluster 3
Gaussian Mixture
LOG AverageFruitLength_cm
LOG AverageFruitWidth_cm
AverageFruitLength_cm
AverageFruitWidth_cm
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00

Table of Content

- Overview
- Objective
- Data Source
- Data Exploration: Palm Subfamily
- Dimensionality Reduction Analysis
- Unsupervised Modelling
- Modelling Evaluation
- Limitation
- Conclusion

Presentation Duration : 20 min

Modelling Limitation

- Number of feature inputs are low. More information about plant species like leaf size shape, fruit color, growth requirement in water, soil, climate etc may help in cluster data points better.

- Highly complex relations among plant species that may be hard to cluster exclusively.

Future Improvements

- Run Modelling on reduced dimensionality components (PCA for example)

- Addressing the missing values by data fill in using supervised learning methods.

Table of Content

- Overview
- Objective
- Data Source
- Data Exploration: Palm Subfamily
- Dimensionality Reduction Analysis
- Unsupervised Modelling
- Modelling Evaluation
- Limitation
- Conclusion

Presentation Duration : 20 min

Conclusion: Part 1

- Cluster analysis on Palm Subfamily shows overall consistent distribution of datapoints.

- Bivariate distribution for each predicted cluster is similar among the models.

Conclusion: Part 2

- The predicted clusters and taxonomic grouping (Subfamily) are not in agreement. This implies that datapoint sharing the same physical traits (fruit properties, stem, growth etc) could potentially originate from different Subfamily.

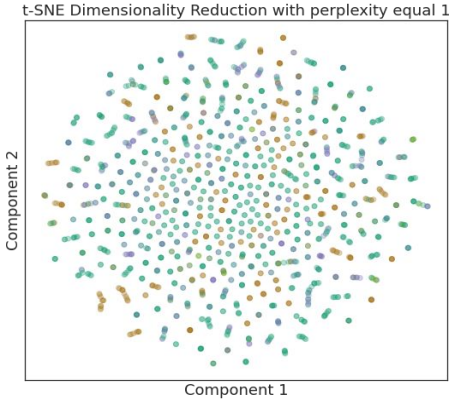
- This information can be used to advantage in plantation planning or forest regeneration of species with shared traits

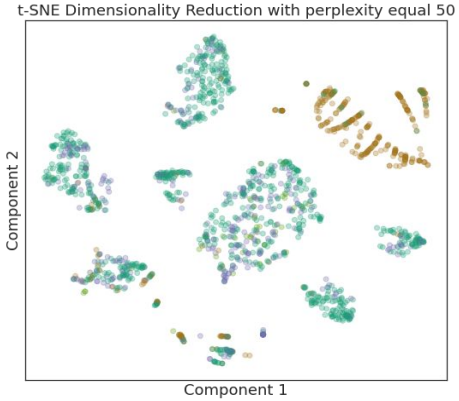
Question ?

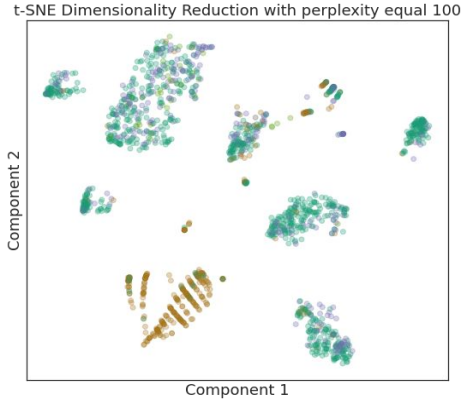
Appendix

Dimensionality Reduction: T-SNE Tuning

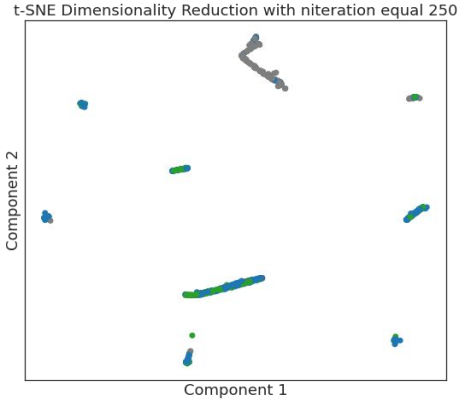
Perplexity: 50

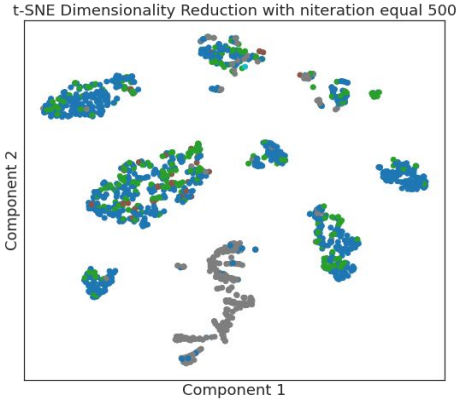


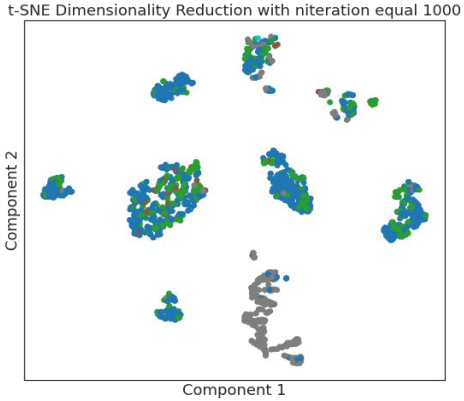




Iteration: 500

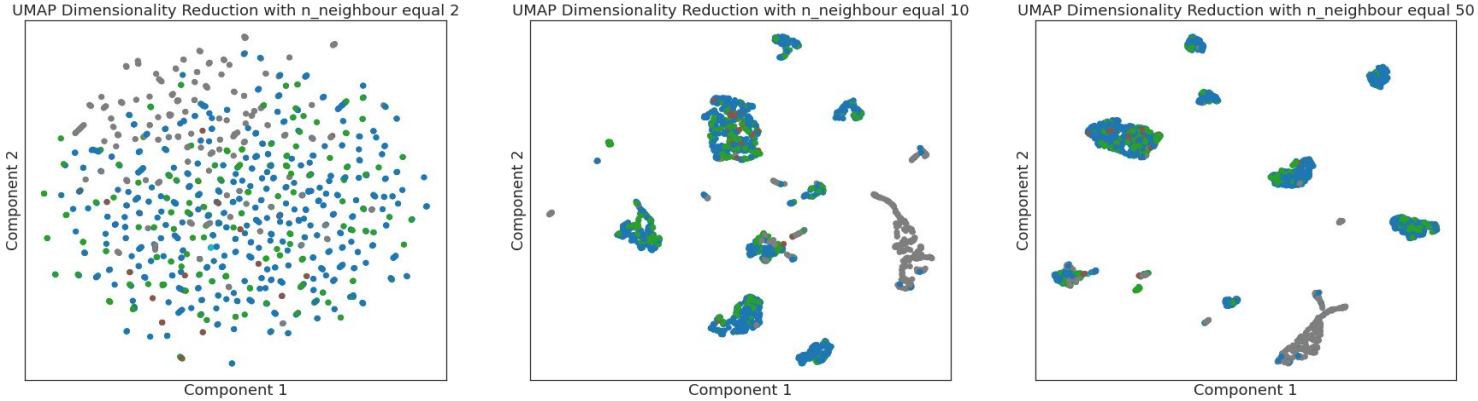




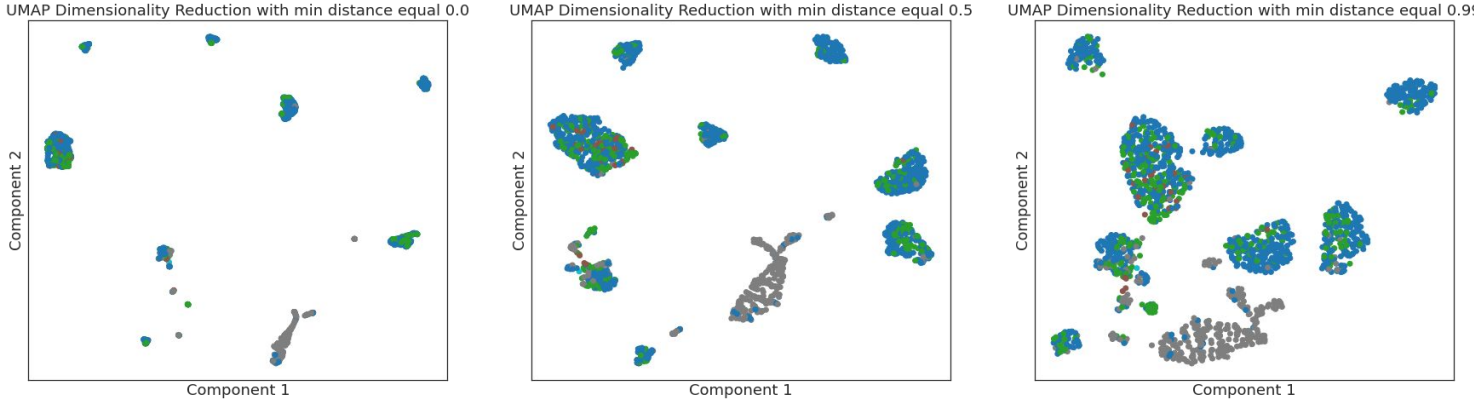


Dimensionality Reduction: U-MAP Tuning

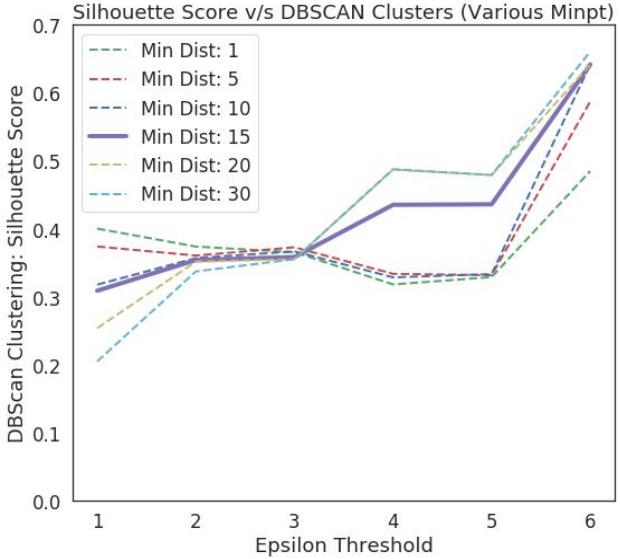
n_ neighbour: 10

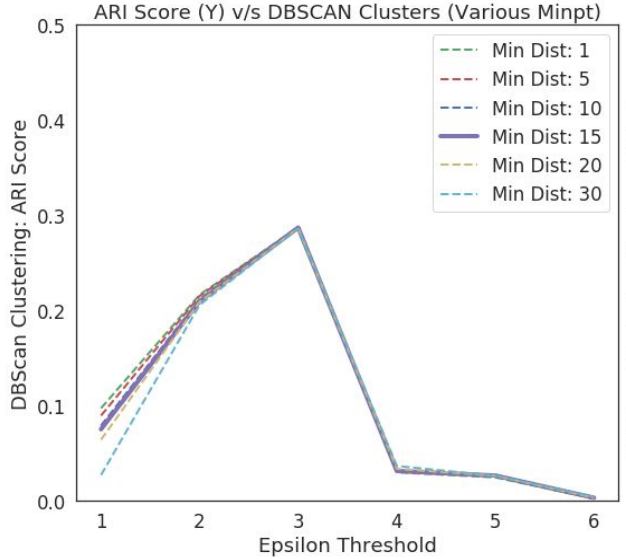


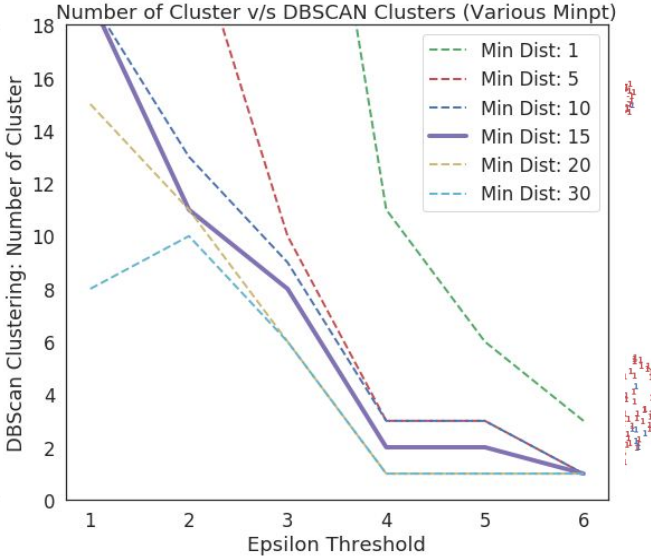
Min distance: 0.5



DB Scan Clustering







- Maximum Silhouette Score: 0.6 for epsilon:6 (max limit, fails to model any clusters)
- ARI Score Y: 0.30 for epsilon: 3