CLASSIFYING THE MS AND IR SPECTRA OF FRUIT USING K-MEANS CLUSTERING AND SUPPORT VECTOR MACHINES

András Vékássy

CHEM6164

X May 2023

Structure

- Exploratory Data Analysis
- Dimensionality Reduction
- Clustering
- Support vector classification
- Conclusion & Future Work
- Q&A

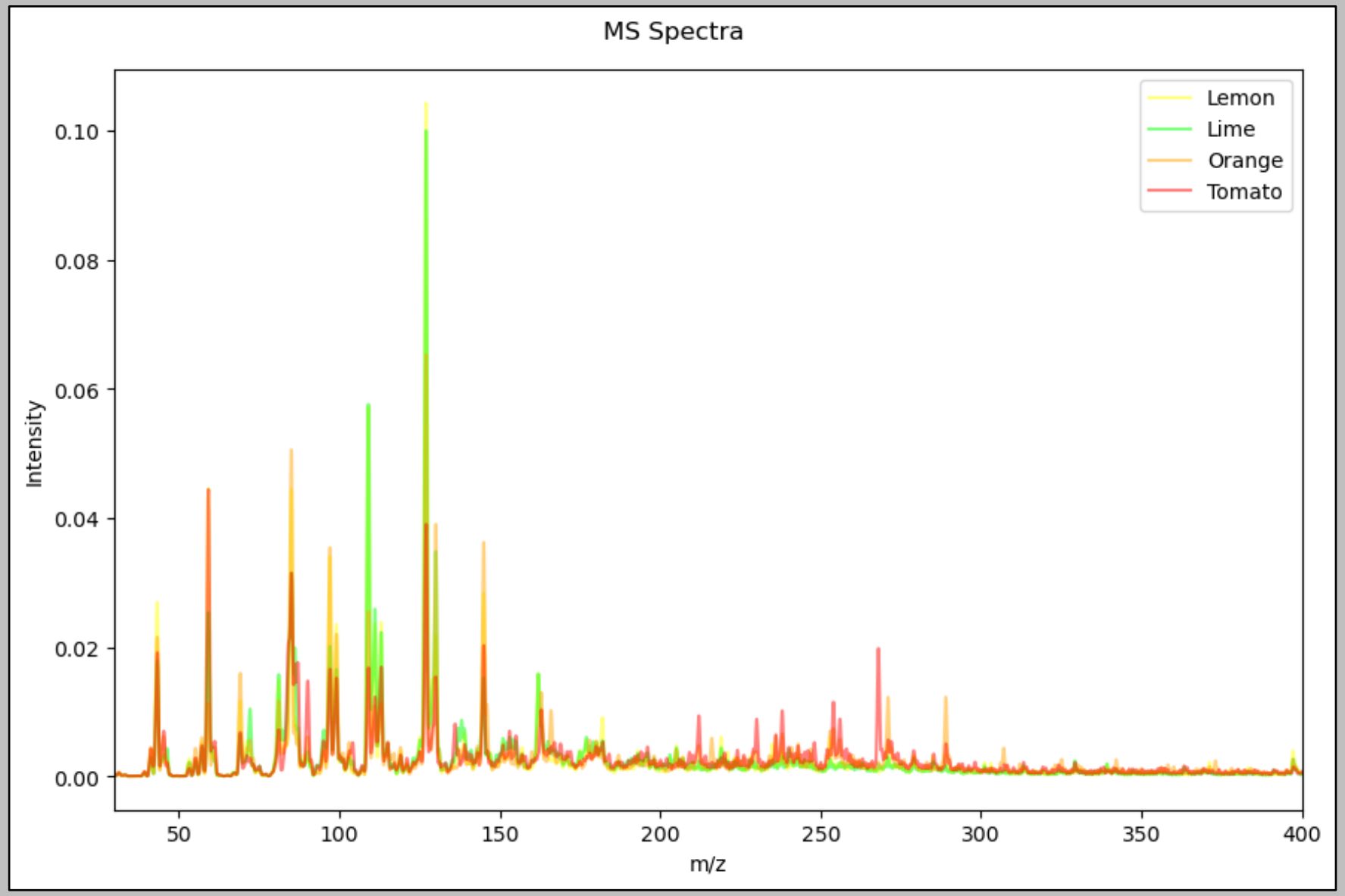
Title

Key message

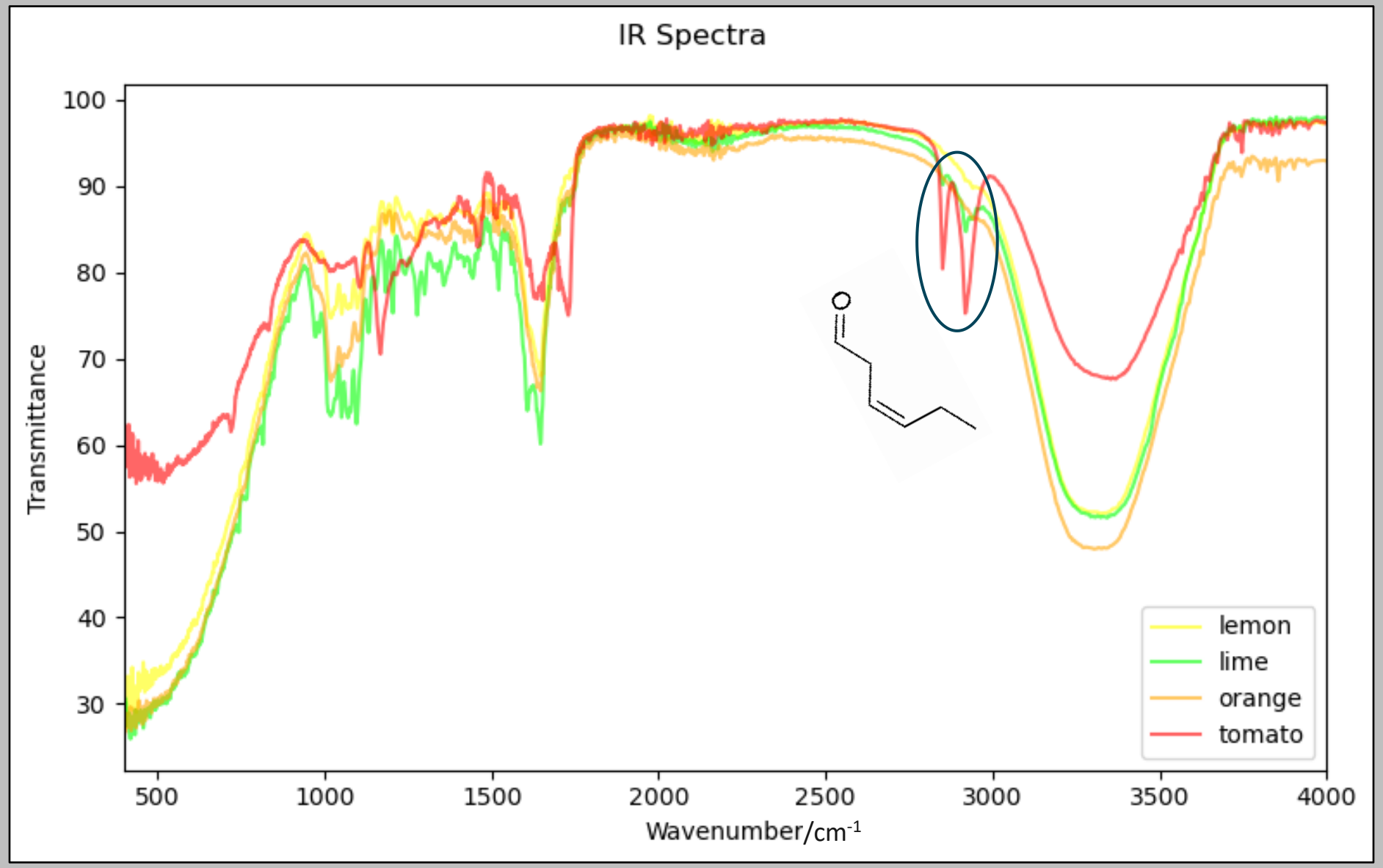
- A

EDA

EDA – MS


MS Spectra
Lemon
Lime
Orange
Tomato
Intensity
0.10
0.08
0.06
0.04
0.02
0.00
50
100
150
200
250
300
350
400
m/z

EDA – IR

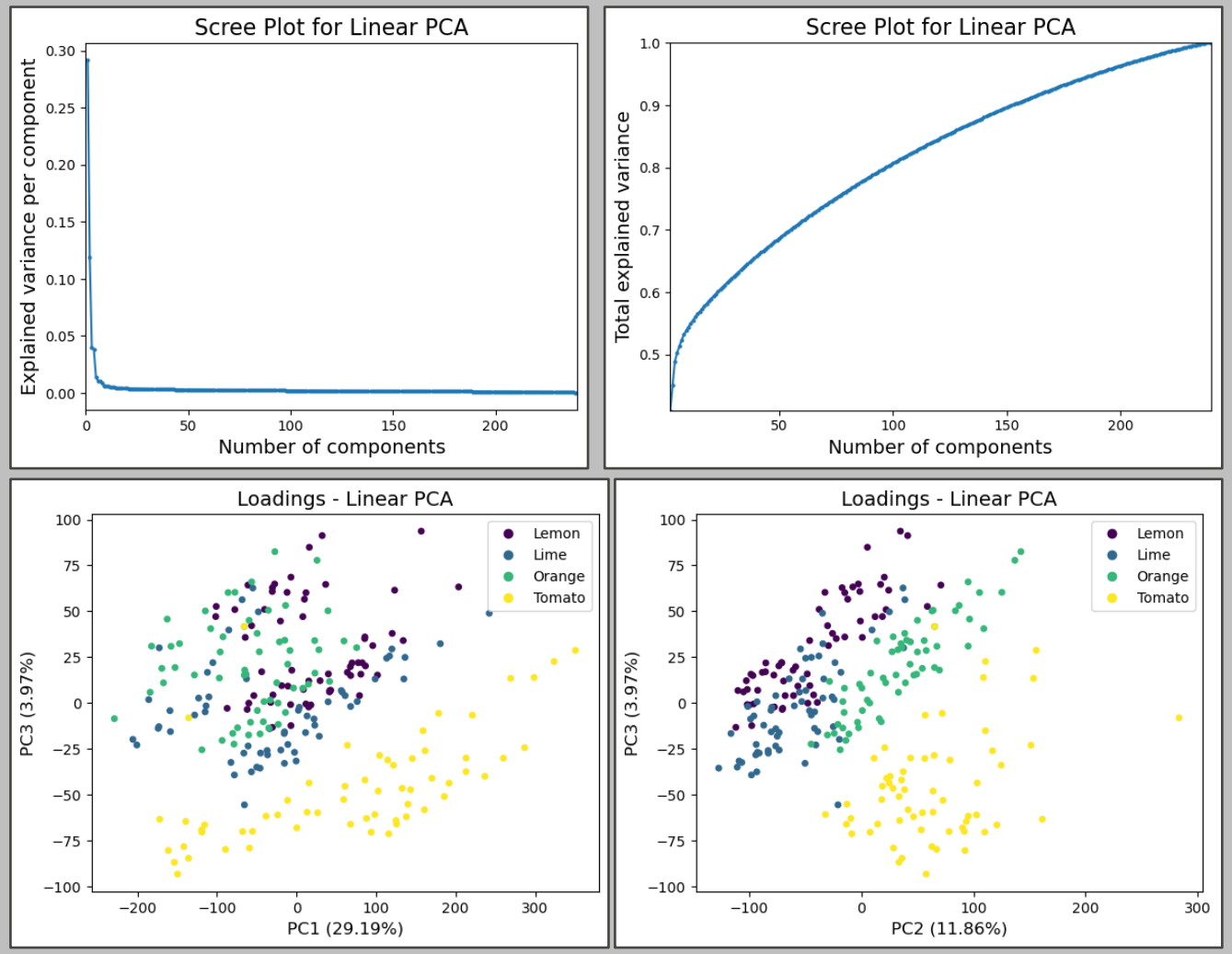

IR Spectra
Transmittance
100
90
80
70
60
50
40
30
500
1000
1500
2000
2500
3000
3500
4000
Wavenumber/cm-1
lemon
lime
orange
tomato

PCA

Dimensionality Reduction – MS

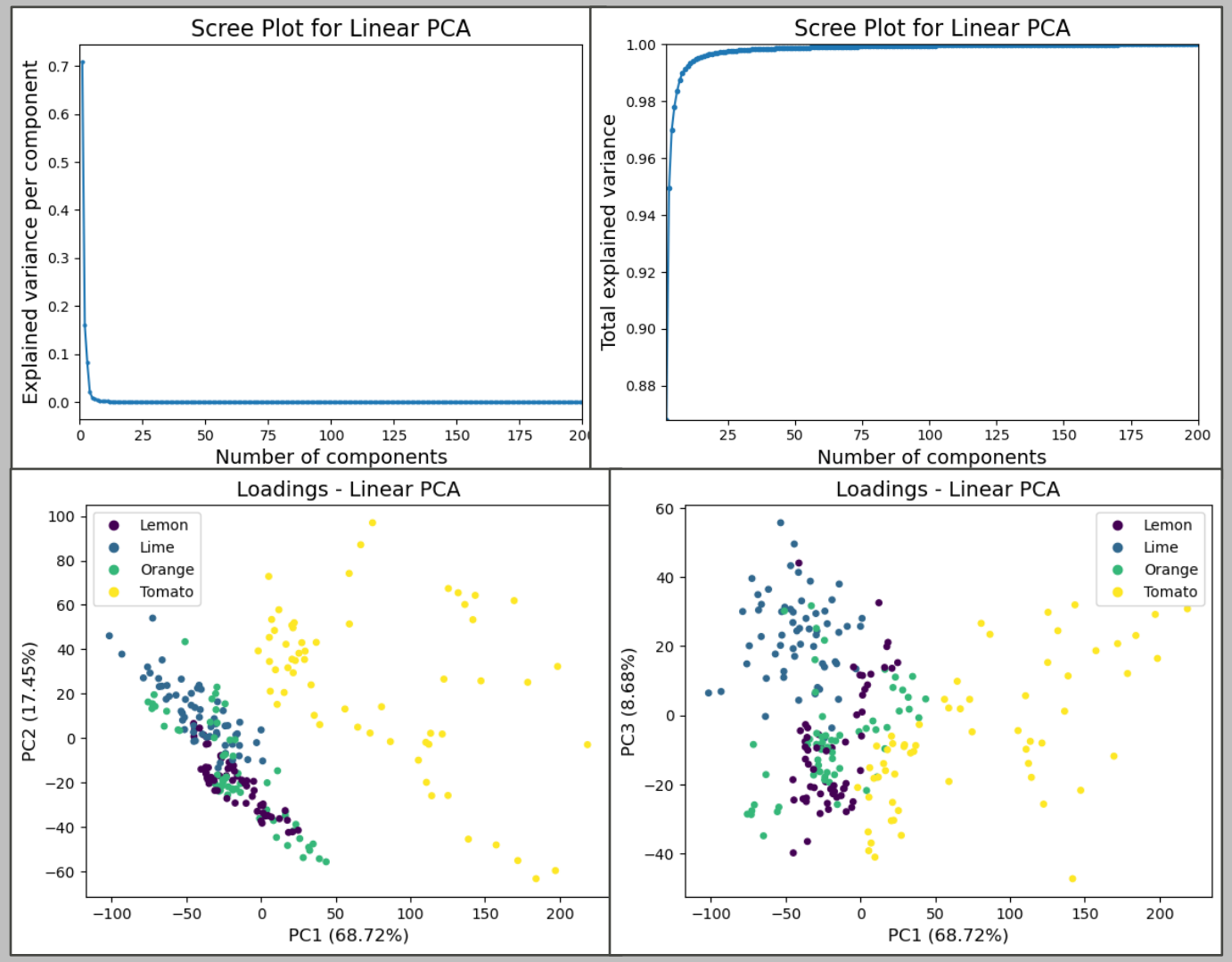
KM

- 90% of variance captured by 150 PCs
- No distinct clusters along PC1
- PC2 and PC3 separated orange from lemon and lime clusters



Significantly sparser dataset

- First 10 PCs capture almost all variance
- PC1 separated tomato from citrus fruits
- PC3 distinguished lime

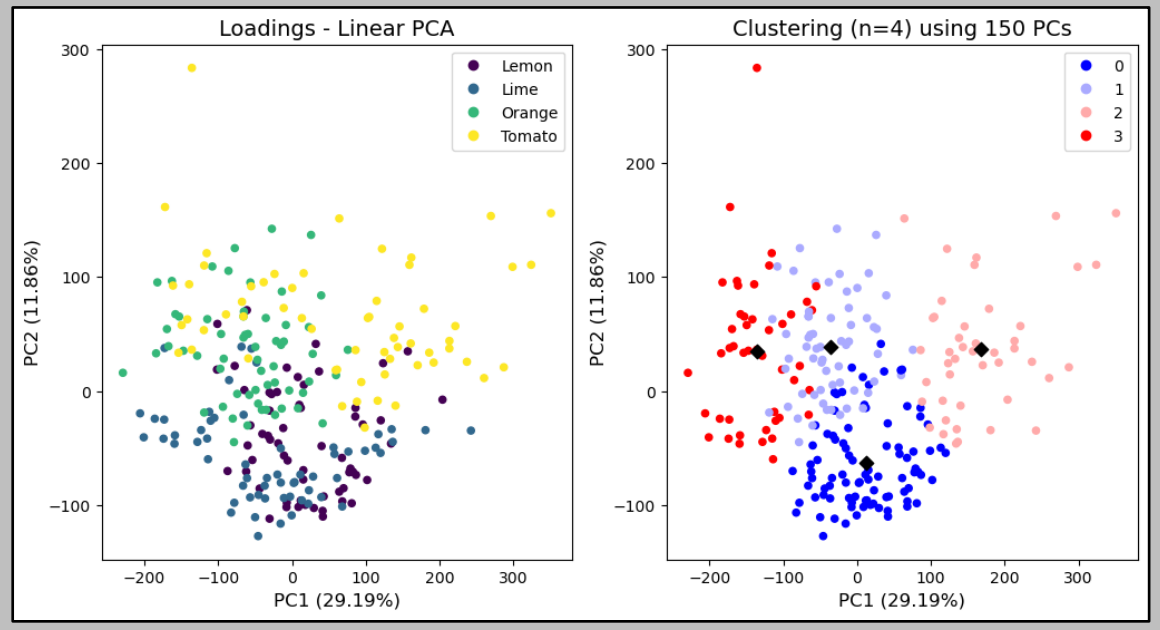


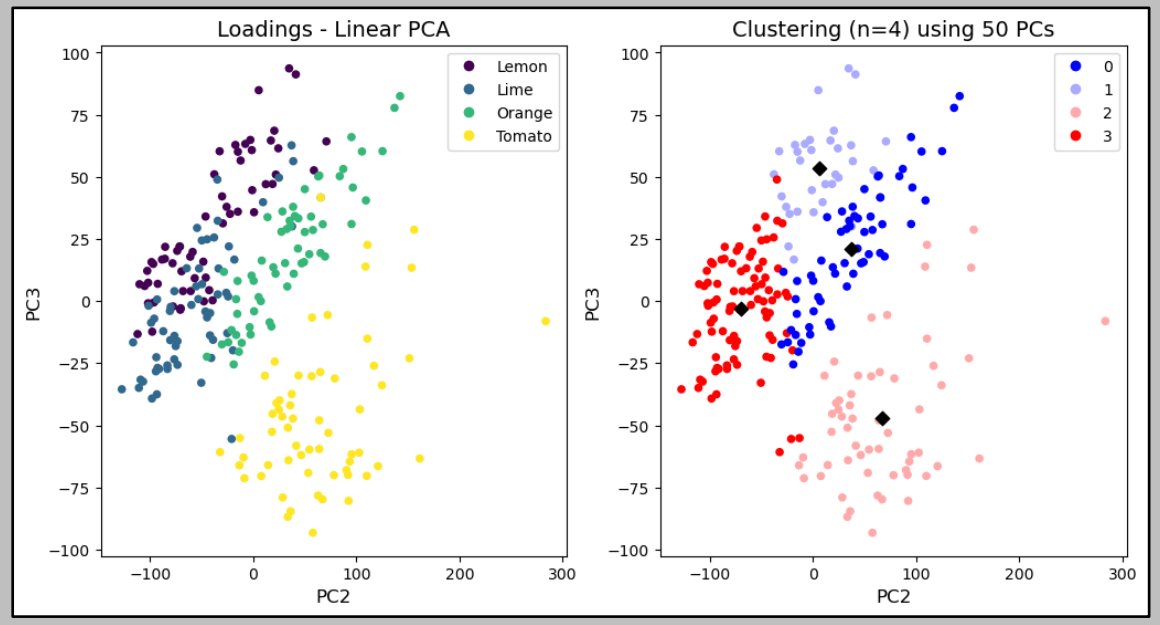
CLUSTERING

Clustering – MS

Not seeing the forest for the trees

- Clustering was misled by PC1
- Number of PCs had no influence on accuracy
- Lime-lemon pair remained unresolved even when using more clusters (n>4)

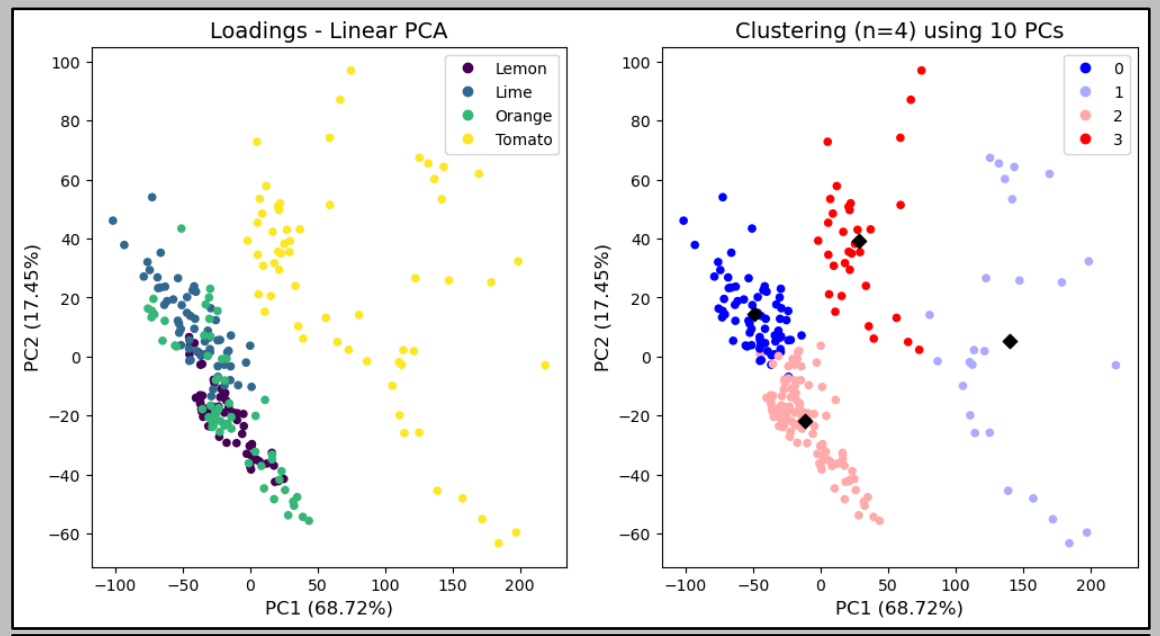


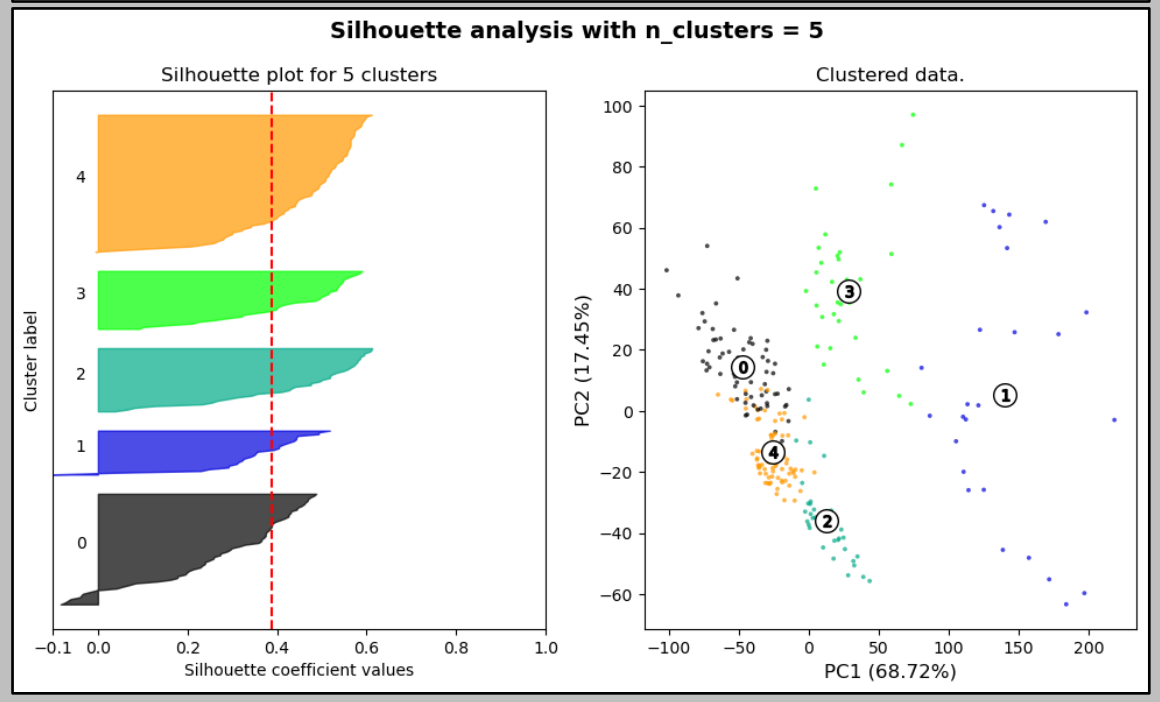


Clustering – IR

The more clusters, the better?

- Silhouette scores increased with less PCs
- Combining clusters is a valid strategy for classifying tomatoes

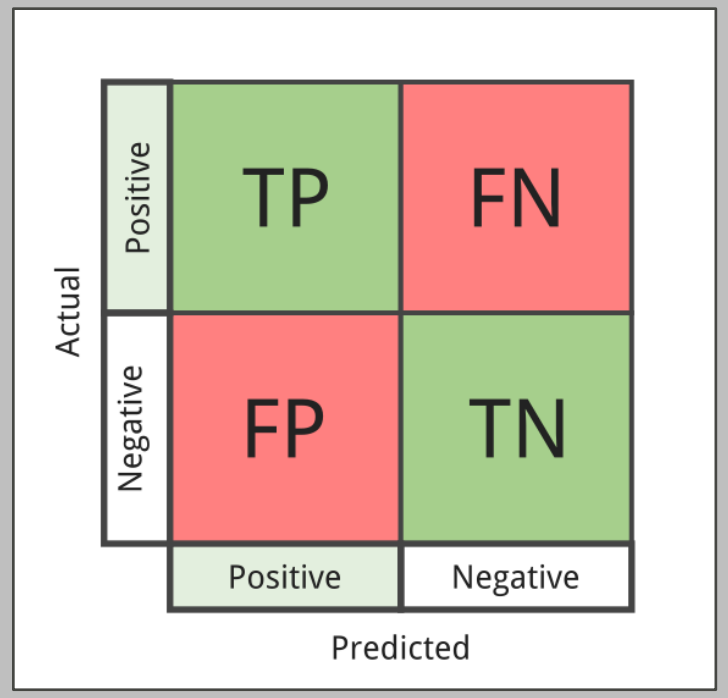




SVM

Success metrics

- Accuracy: (TP+TN)/A
- F1 score: (2 x precision x recall)/(precision+recall)
- Precision: TP/(TP+FP)
- Recall: TP/(TP+FN)
- Confusion matrix



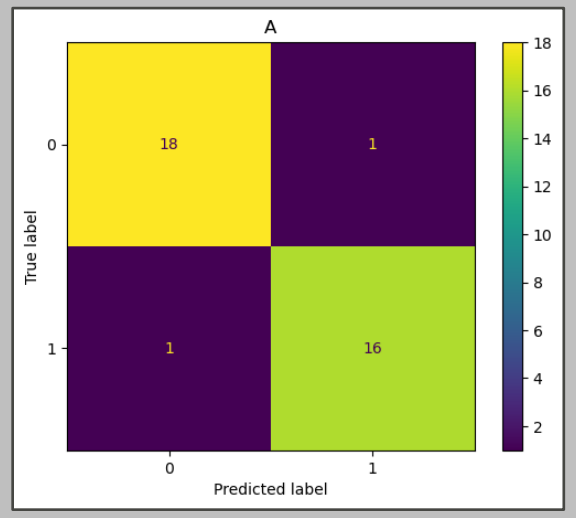
Key message

Separable pair: orange-tomato

- Accuracy and F1 scores all 1.0
- RBF failed to classify

Non-separable pair: lime-lemon

- Great separation using C=0.01
- Varying the number of PCs lowered recall



#PCs	Accuracy	F1 score	Precision	Recall
2.0	0.472222	0.000000	0.000000	0.000000
3.0	0.694444	0.645161	0.833333	0.526316
5.0	0.916667	0.923077	0.900000	0.947368
10.0	0.972222	0.972973	1.000000	0.947368
20.0	1.000000	1.000000	1.000000	1.000000
50.0	0.916667	0.918919	0.944444	0.894737
100.0	0.916667	0.923077	0.900000	0.947368

The confusion matrix of the SVM model using C=0.01 for distinguishing limes and lemons (left) and the SVM's dependence on the number of PCs (right).

SVM – IR

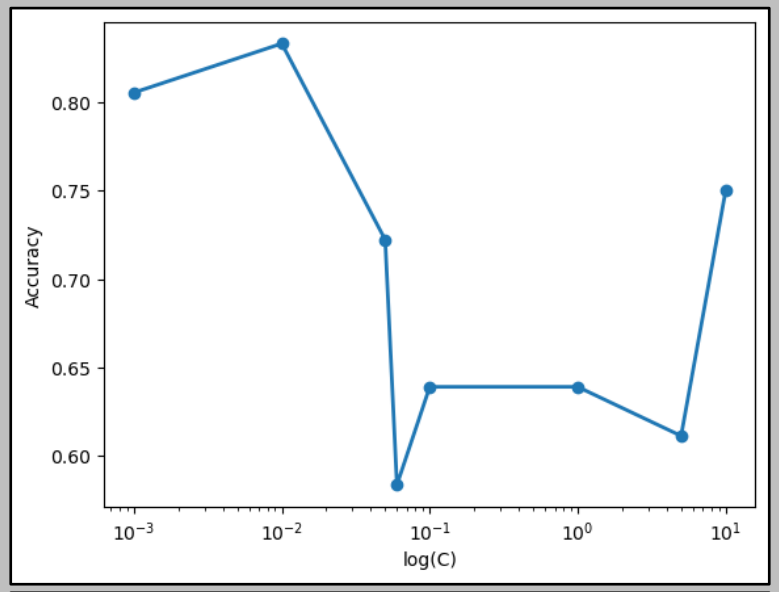
PCs matter

Separable pair: orange-tomato

- Accuracy and F1 scores all 1.0
- RBF made a better attempt

Non-separable pair: lemon-orange

- Softer margin improved results
- Large dependence on the number of PCs



#PCs	Accuracy	F1 score	Precision	Recall
1.0	0.416667	0.400000	0.411765	0.388889
2.0	0.611111	0.681818	0.576923	0.833333
3.0	0.583333	0.705882	0.545455	1.000000
4.0	0.527778	0.679245	0.514286	1.000000
5.0	0.555556	0.692308	0.529412	1.000000
6.0	0.555556	0.692308	0.529412	1.000000
7.0	0.555556	0.692308	0.529412	1.000000
8.0	0.555556	0.692308	0.529412	1.000000
9.0	0.750000	0.800000	0.666667	1.000000
10.0	0.833333	0.857143	0.750000	1.000000

CONCLUSION & FUTURE WORK

When life gives you lemons…

- PCA reduced dimensionality by up to two magnitudes
- k-means clustering struggled with overlapping data
- SVMs tackled non-separable classes very well

Next steps:

- Finetune SVM's parameters with gradient search
- Try alternative algorithms: tree-based models, neural networks (CNNs, transfer learning)
- *Explain* the models

REFERENCES

References

(1) R. Houhou and T. Bocklitz, *Anal. Sci. Adv.*, 2021, **2**, 128–141.

(2) N. S. Eyke, B. A. Koscher and K. F. Jensen, *Trends Chem.*, 2021, **3**, 120–132.

(3) M. D. Wilkinson et al. *Sci. Data*, 2016, **3**, 1–9.

(4) Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*, O'Reilly Media, Inc., 2019.

(5) C. A. Peña-Solórzano, D. W. Albrecht, R. B. Bassed, M. D. Burke and M. R. Dimmock, *Forensic Sci. Int.*, DOI:10.1016/j.forsciint.2020.110538.

(6) Y. Alkhalifah et al. *Anal. Chem.*, 2020, **92**, 2937–2945.

(7) R. Houhou and T. Bocklitz, *Anal. Sci. Adv.*, 2021, **2**, 128–141.

(8) Y. Pérez, M. Casado, D. Raldúa, E. Prats, B. Piña, R. Tauler, I. Alfonso and F. Puig-Castellví, *Anal. Bioanal. Chem.*, 2020, **412**, 5695–5706.

(9) L. Xu, Y. Liu, J. Yu, X. Li, X. Yu, H. Cheng and J. Li, *J. Neurosci. Methods*, 2020, **331**, 108538.

(10) K. Weiss, T. M. Khoshgoftaar and D. Wang, *J. Big Data*, 2016, **3**, 9.

(11) R. Guidotti, A. et al. *ACM Comput. Surv.*, DOI:10.1145/3236009.

(12) R. G. Buttery, R. Teranishi and L. C. Ling, *J. Agric. Food Chem.*, 1987, **35**, 540–544.

(13) C. Molnar, *Interpretable Machine Learning*, 2nd edn., 2022.

(14) S. M. Lundberg, G. G. Erion and S.-I. Lee, *arXiv Prepr. arXiv1802.03888*.

Q&A

Title

Key message

- A

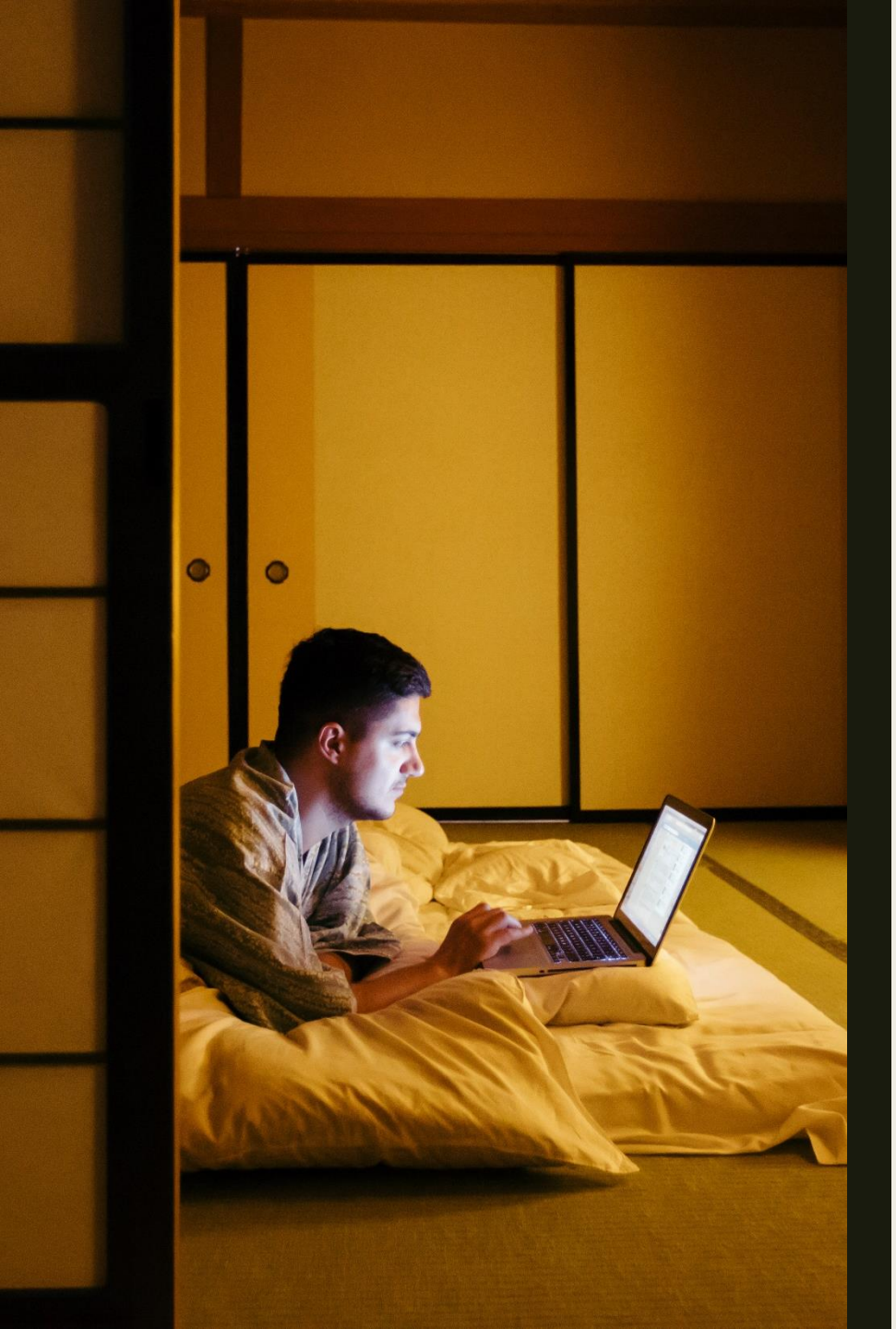
CONCLUSION & FUTURE WORK

QSAR

TRAVEL DESIGN

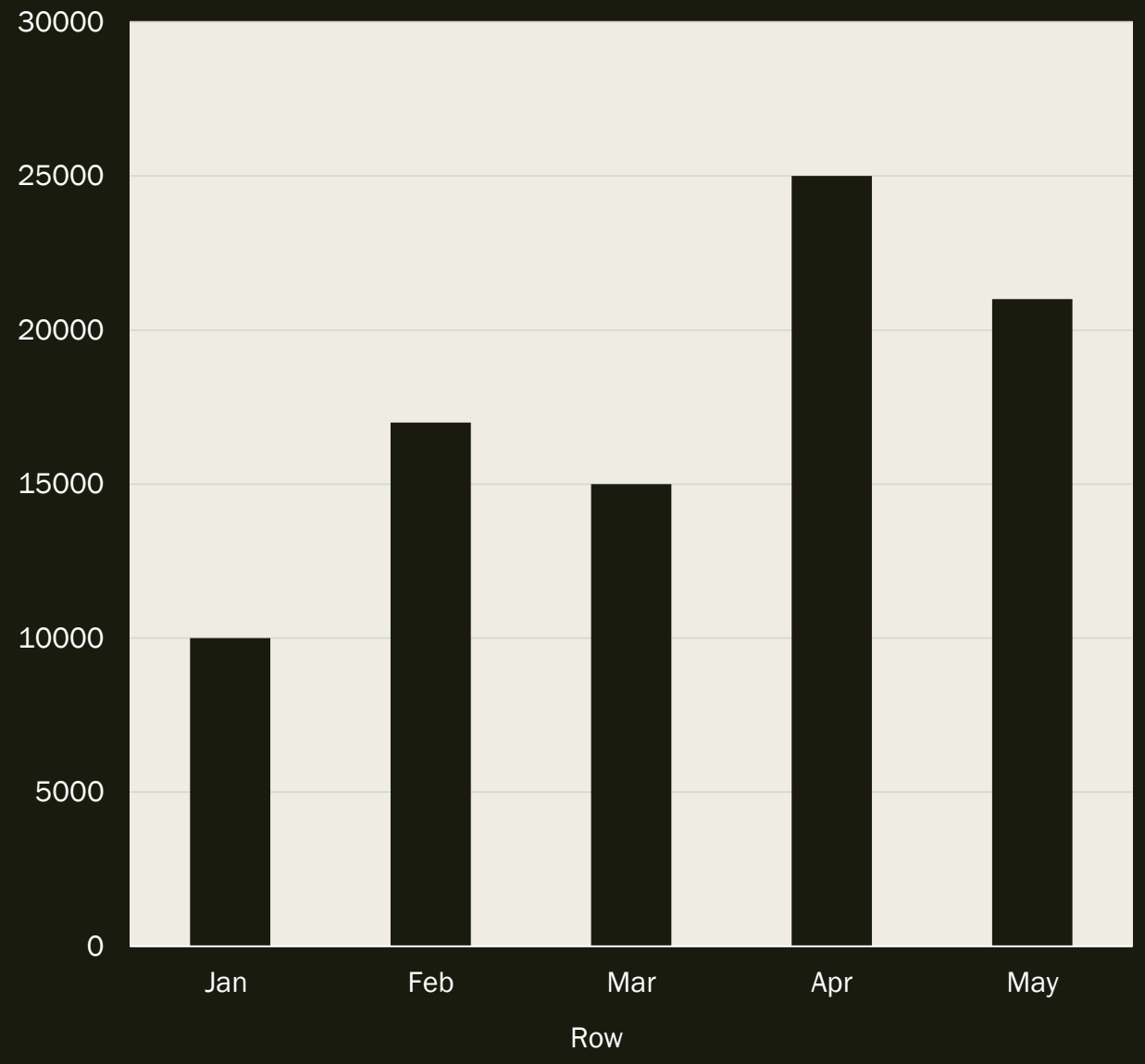
Digital Nomad Guide:

The better way to enjoy your world as a nomad



How to Start Being a Digital Nomad

- Reduce and Eliminate Expenses
- Decide on a Location
- Set Goals and Create a Plan



Common Travel Seasons

Make a Plan

1. Set a destination and/or travel theme
2. Set travel dates, mode of travel, and find accommodations
3. Find local attractions, food venues, and activities

THANK YOU

someone@example.com