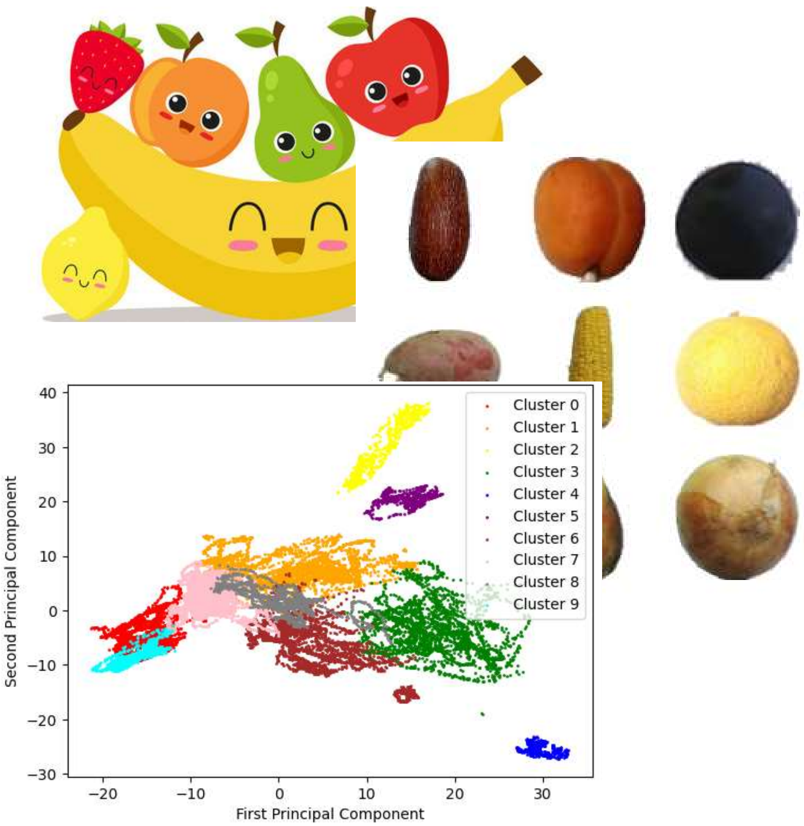
Fruit Image Exploration:
Dimension Reduction and Clustering Study

Nick Vastine
DTSA5510 – Unsupervised Algorithms in Machine Learning
Final Project
August 19, 2024

Data Science
UNIVERSITY OF COLORADO **BOULDER**

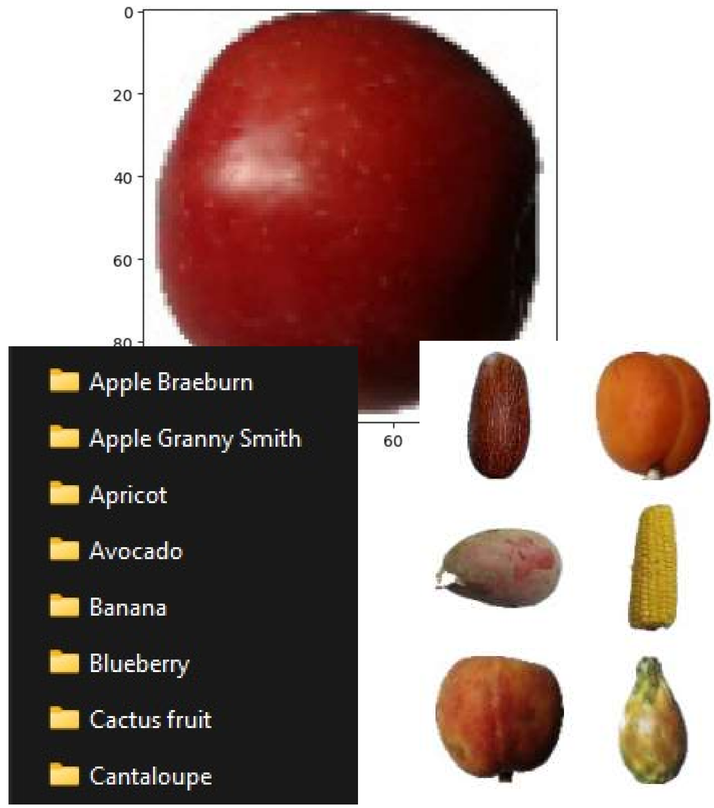
Project Topic

- Perform unsupervised classification of fruit images.
- Explore the effects of feature reduction, both manually and via PCA to create two models.
- Compare clustering methods and perform hyperparameter tuning on the number of clusters.



Data / Data Cleaning

- Data is sourced from Kaggle
- 22,495 Sample Images
  - Each image is 100 pixels x 100 pixels
  - Each pixel is assigned an R,G, and B value from 0-255
  - This results in a (22495,100,100,3) sample matrix
- Each image was opened and read into the central matrix



Exploratory Data Analysis

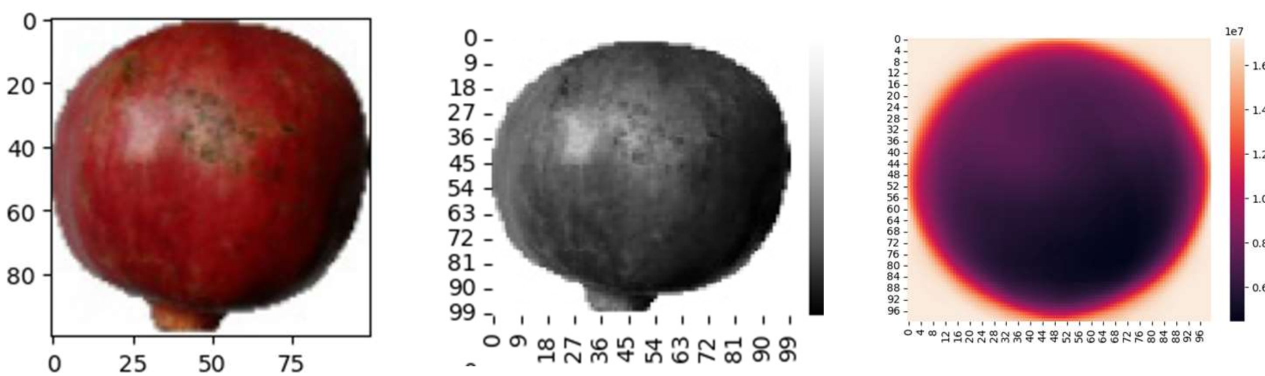
- **Understanding Features**
  - There are two relationships within our data
    1. *Color value* of each pixel, and therefore color of the object

    2. *Shape* of the object

  - We could simply flatten (100,100,3) features into *30,000 features.* But that’s a lot of information to digest…

Exploratory Data Analysis

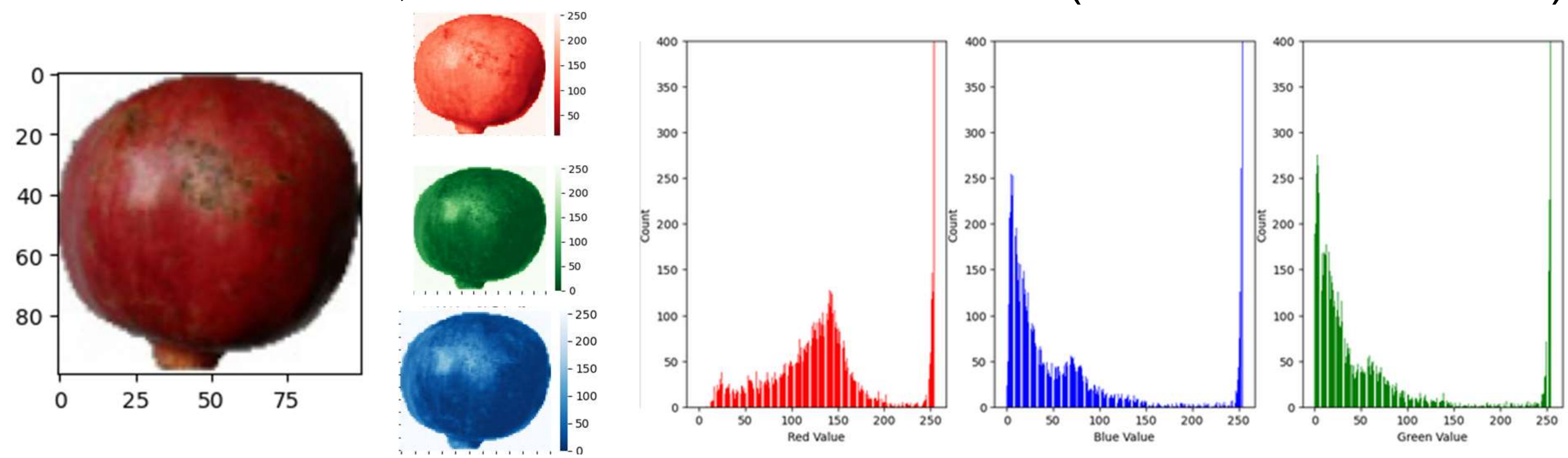
- ***Model 1*** *- Intensity / Greyscale*
    - Consider if we average the RGB values into a single '*intensity'* value.
    - This reduces 30,000 features to *only* 10,000 features!



- Note however this effectively disregards our color information…

Exploratory Data Analysis

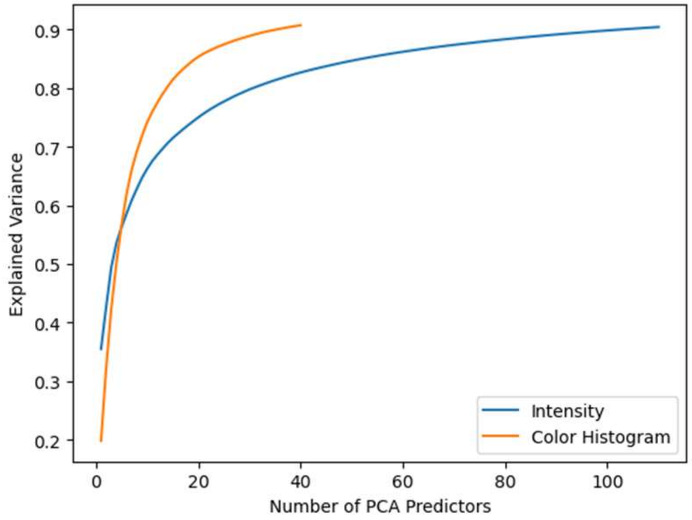
- ***Model 2*** *– Color Histogram*
    - Consider if take histograms of each R, G, and B value, then store those histograms as our sample feature matrix.
    - This reduces 30,000 features to *765* features! (3 colors * 255 values).



- Note however this disregards our relative position information...

Modeling – Flattening, Scaling, PCA

- First, flatten the data into the *Intensity matrix* (22495, 10000) and process our *Color Histogram matrix* (22495, 765).
- This data is then scaled before we perform PCA.



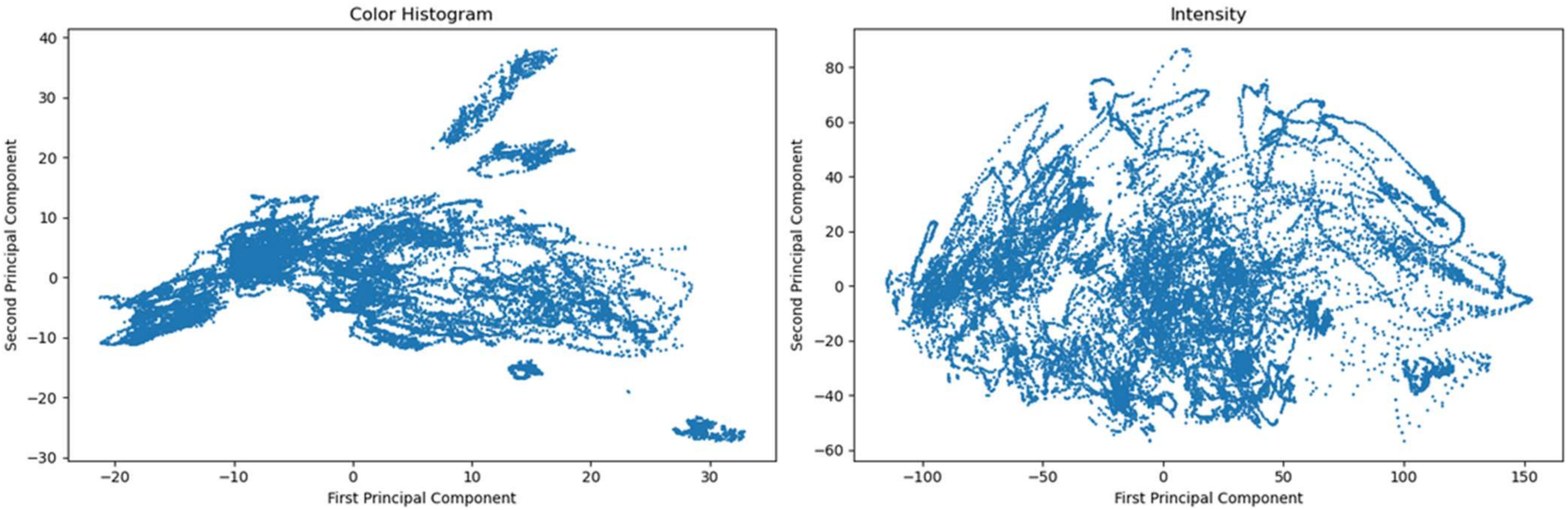
- The PCA code determined the number of principal components to explain 90% of the variance.

```
pca_intensity_mod = PCA(n_components = 0.9, svd_solver = 'full').fit(intensity_data_scaled)
pca_color_hist_mod = PCA(n_components = 0.9, svd_solver = 'full').fit(color_hist_data_scaled)
```

- This analysis found the *Color Histogram* model required 40 components, while *Intensity* required more than 100.

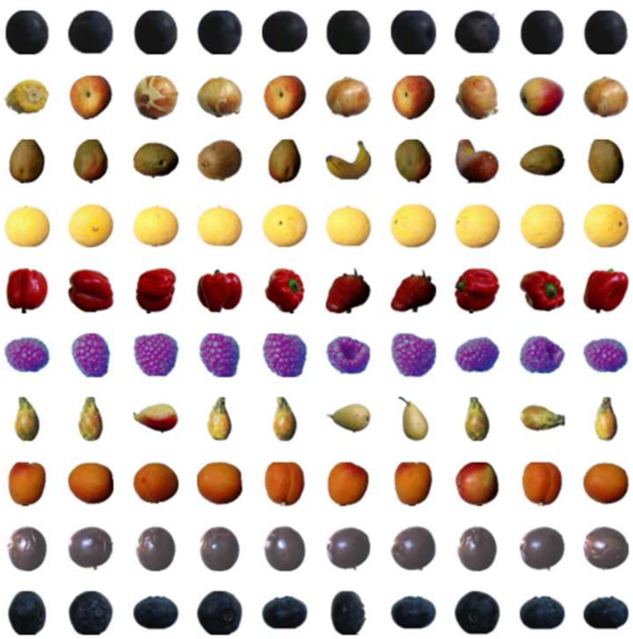
Modeling – PCA (cont.)

- We can also visualize how the first two principal components work to separate our data into possible clusters for each model.

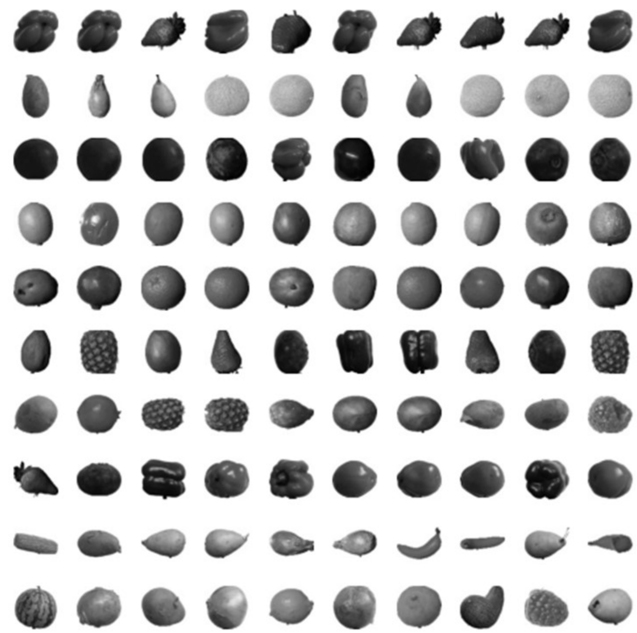


Modeling – Reduction Comparison

- We will first evaluate arbitrarily using 10 clusters.



Color Histogram Sample



Intensity Sample

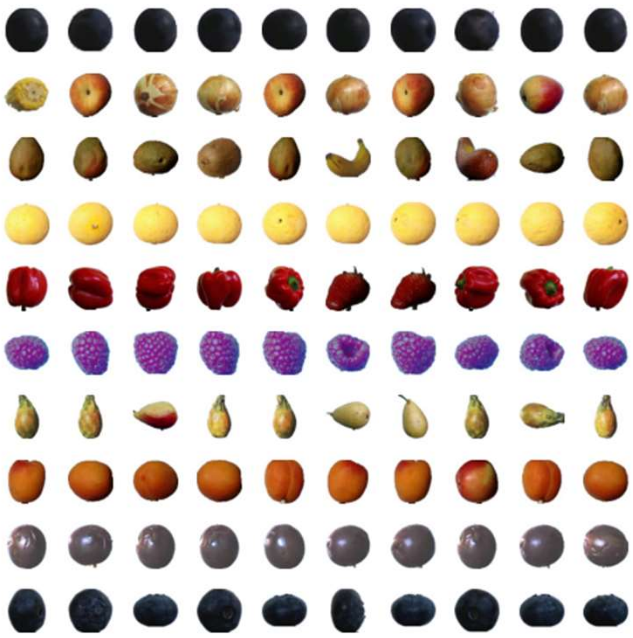
Modeling – 33 Clusters

- We know there are 33 categories from the training data.

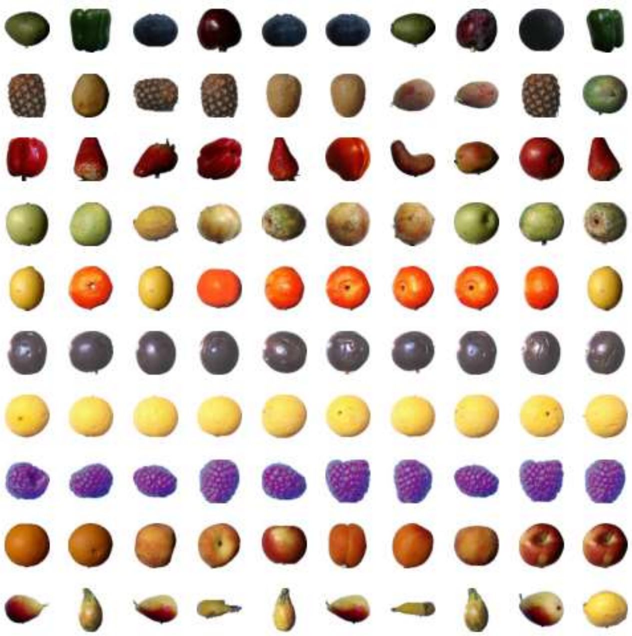
33 Cluster Color Histogram Sample

Modeling – Method Comparison

- We can compare methods for our color histogram dataset



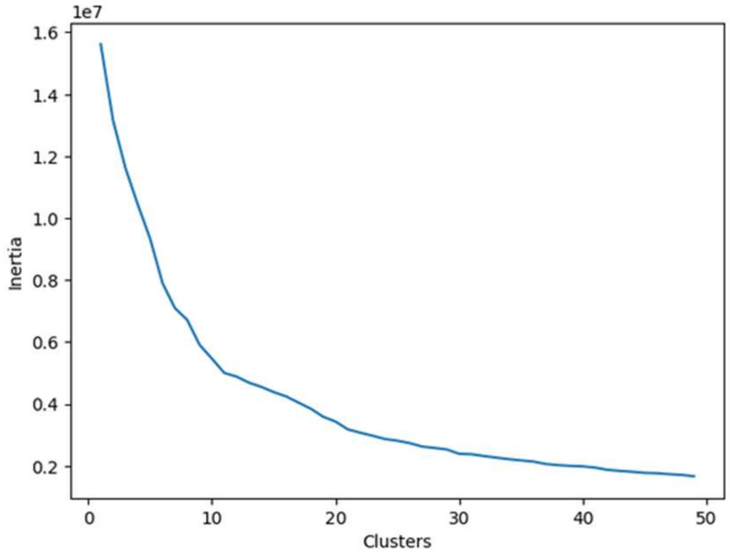
K-Means Sample

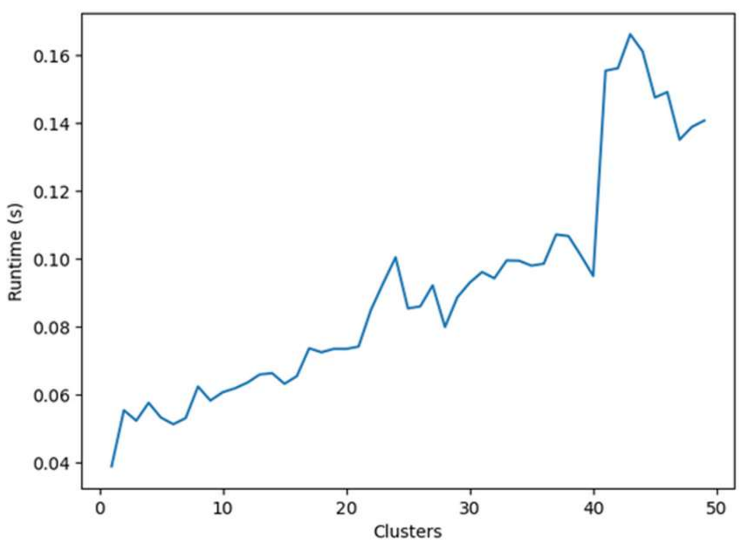


Agglomerative Sample

Modeling – Hyperparameter Tuning

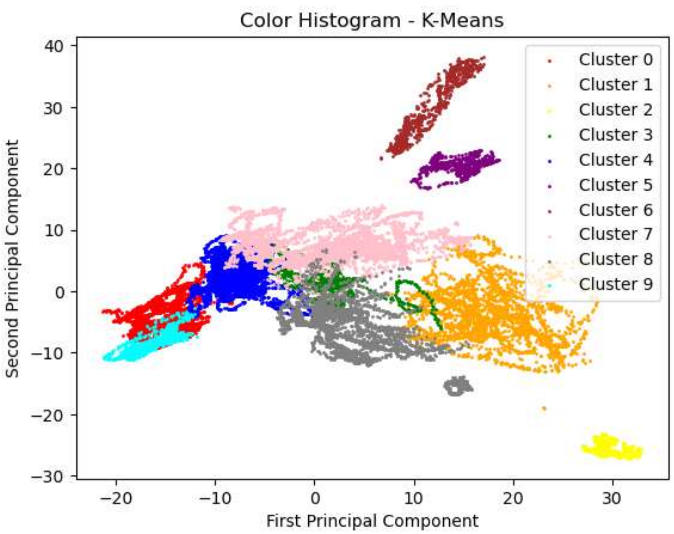
- Below evaluates our color histogram K-means across a range of clusters, comparing inertia and runtime.
  - Note the ‘elbow’ at 10 clusters, conveniently matching our analysis.

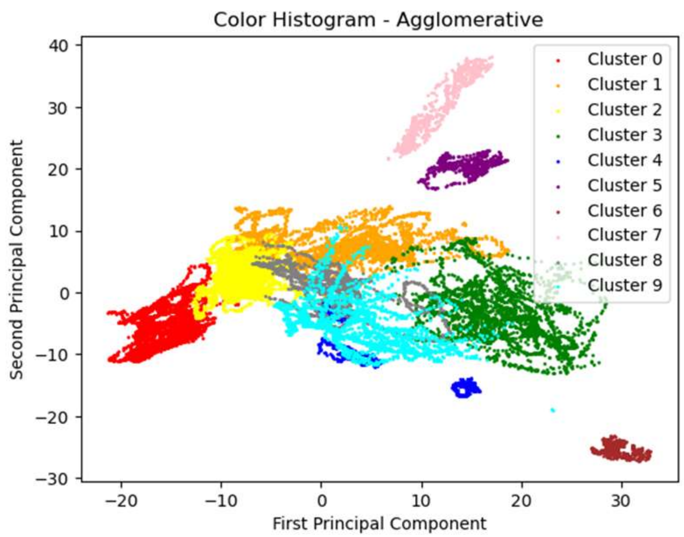


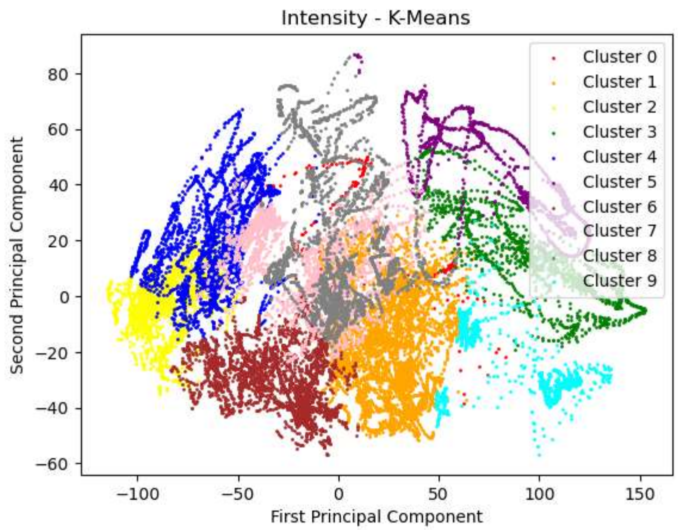


Modeling – Method Comparison

- We can compare based on our principal components as well







Results & Analysis Summary

- Dimension Reduction is inherently a sacrifice in resolution
  - Using a mix of approaches can dramatically reduce features (from 30,000 to 765 to 40 for the color histogram)
  - There are tradeoffs (ex. loss of color information for intensity analysis)

- It is valuable to visualize and contextualize your response
  - In this case, an imagery problem makes results easy to interpret.
  - PCA analysis plots are valuable to anticipate clusters, or if data is suitable for a clustering application

- K-Means runs *significantly* faster than Agglomerative Clustering

Discussion & Conclusion

- Key takeaway is the importance of *experimentation, appropriate feature reduction*, and *evaluation procedures*.

- There is such thing as *too much data* (curse of dimensionality)
  - Was unable to perform PCA on the flattened original dataset

- Future Improvements
  - Structure the analysis for supervised learning, using the labeled training set to calculate accuracy for 33 categories
  - Consider applying *IncrementalPCA* when there are too many features
  - Explore the effects of reduced datasets on clustering effectiveness

Thank you for listening!

Fruit Image Exploration:
Dimension Reduction and Clustering Study

Nick Vastine
DTSA5510 – Unsupervised Algorithms in Machine Learning
Final Project
August 19, 2024

Data Science
UNIVERSITY OF COLORADO **BOULDER**