Capstone project

Hitting the top notes

Modeling on fragrance notes to classify ratings

Allison Bishop

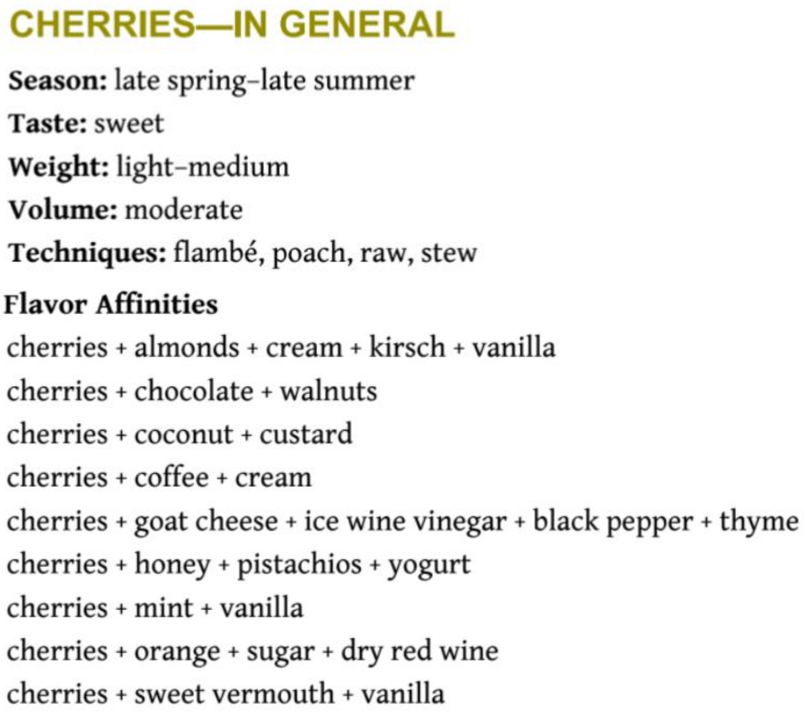
Image source: @_positivity on @fruitassembly IG, hiclipart.com/

Problem statement

Are fragrance notes good predictors of average customer ratings?

Image source: prada.com

Background - My motivation



CHERRIES—IN GENERAL

Season: late spring–late summer

Taste: sweet

Weight: light–medium

Volume: moderate

Techniques: flambé, poach, raw, stew

Flavor Affinities

cherries + almonds + cream + kirsch + vanilla

cherries + chocolate + walnuts

cherries + coconut + custard

cherries + coffee + cream

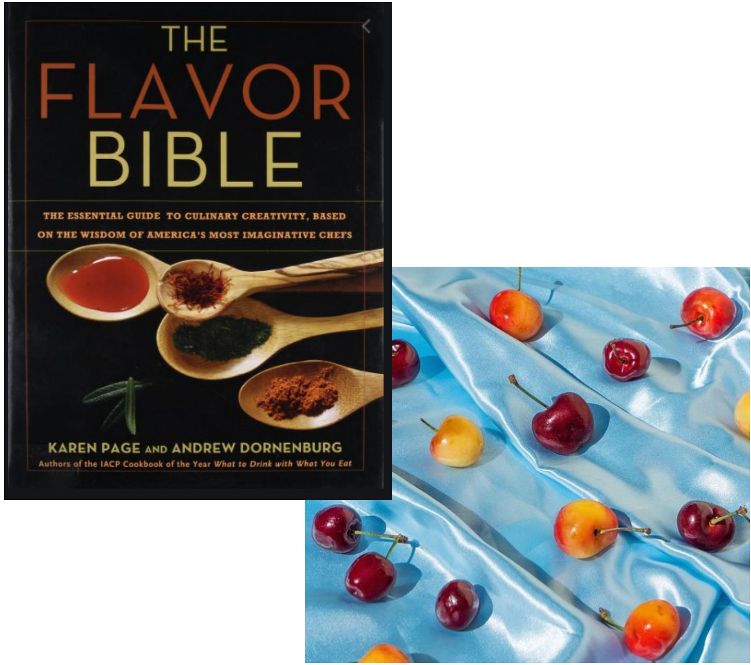
cherries + goat cheese + ice wine vinegar + black pepper + thyme

cherries + honey + pistachios + yogurt

cherries + mint + vanilla

cherries + orange + sugar + dry red wine

cherries + sweet vermouth + vanilla


THE FLAVOR BIBLE
THE ESSENTIAL GUIDE TO CULINARY CREATIVITY, BASED ON THE WISDOM OF AMERICA'S MOST IMAGINATIVE CHEFS
KAREN PAGE AND ANDREW DORNENBURG
Authors of the IACP Cookbook of the Year What to Drink with What You Eat

Image/ data sources: *The Flavor Bible*, Karen Page, Andrew Dornenburg; @morgan.light on @fruitassembly IG

Background - Classes of notes

Top: Form initial impression. Selling point.
High volatility.
Light, bright (like citrus fruits)

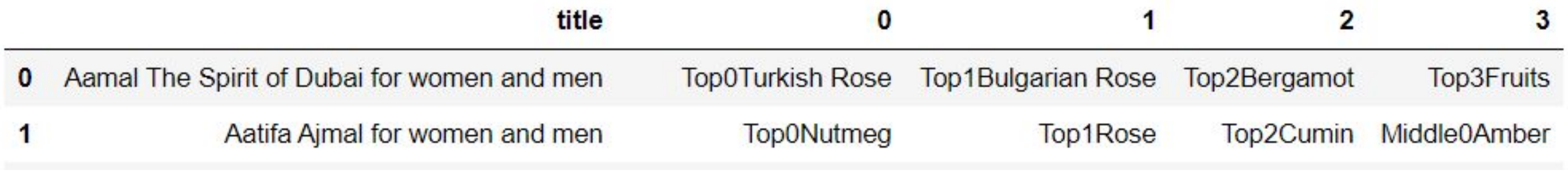
Middle: Forms the body. 40-80% of total aroma.
Midrange volatility.
Complex, midweight (like florals)

Base: Foundation of fragrance. Brings depth.
Low volatility.
Deep, heavyweight (like sandalwood)

Datasource: https://www.alphaaromatics.com/blog/understanding-the-top-notes-heart-notes-base-notes-of-fragrance/

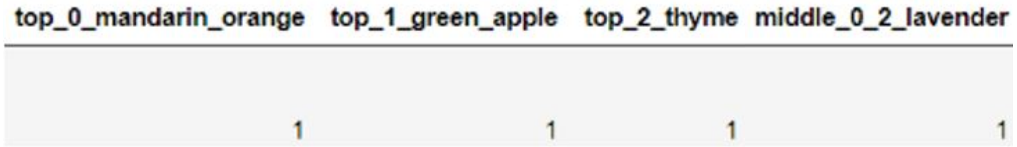
Preprocessing - Dummify notes

Before



	title	0	1	2	3
0	Aamal The Spirit of Dubai for women and men	Top0Turkish Rose	Top1Bulgarian Rose	Top2Bergamot	Top3Fruits
1	Aatifa Ajmal for women and men	Top0Nutmeg	Top1Rose	Top2Cumin	Middle0Amber

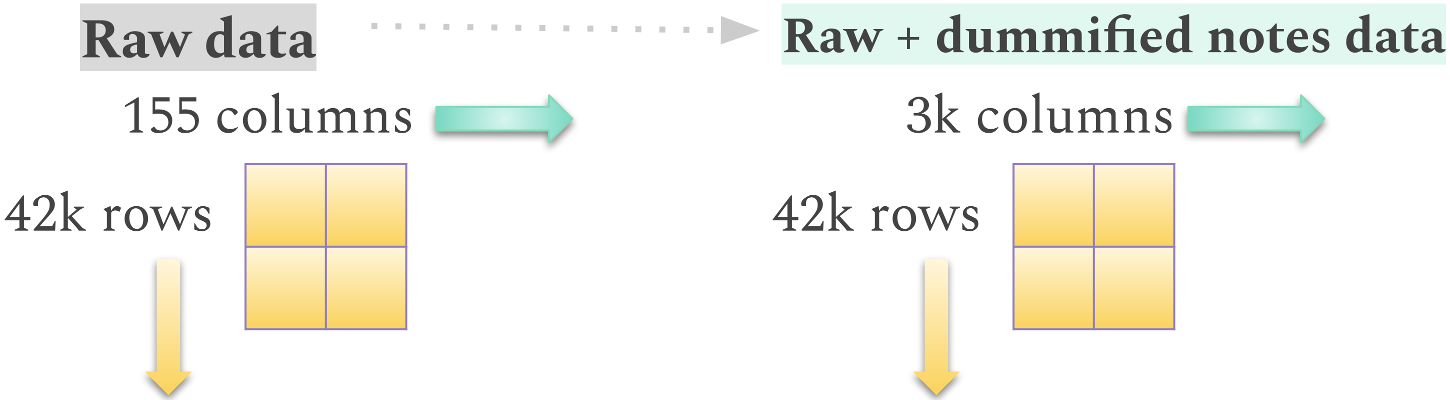
After



top_0_mandarin_orange	top_1_green_apple	top_2_thyme	middle_0_2_lavender
1	1	1	1

Preprocessing - Results

The dataset drastically changed shape



Raw data

155 columns

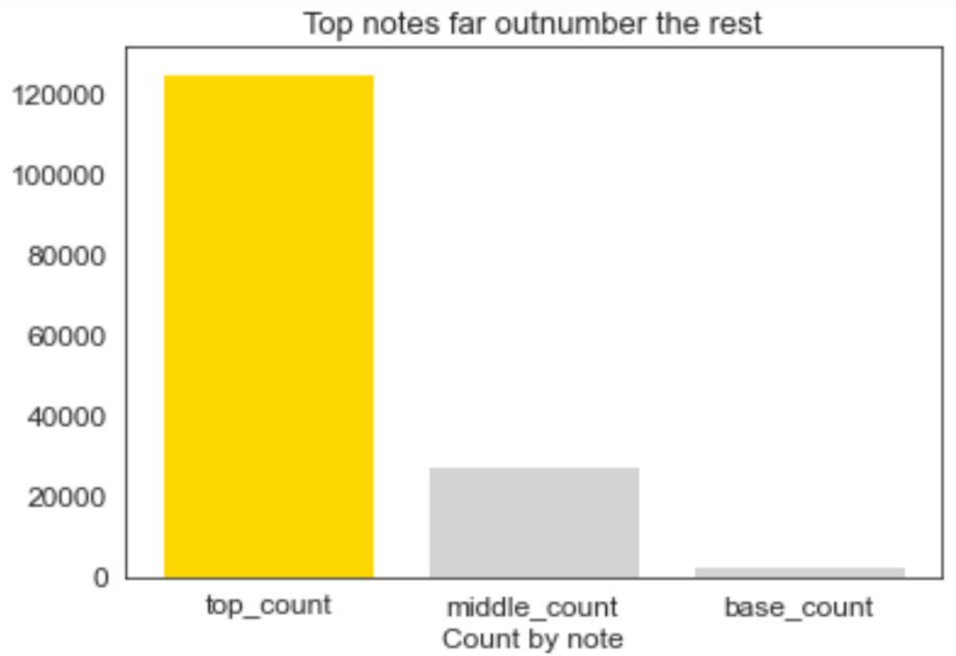
42k rows

Raw + dummified notes data

3k columns

42k rows

Data profile - Top, middle, base notes



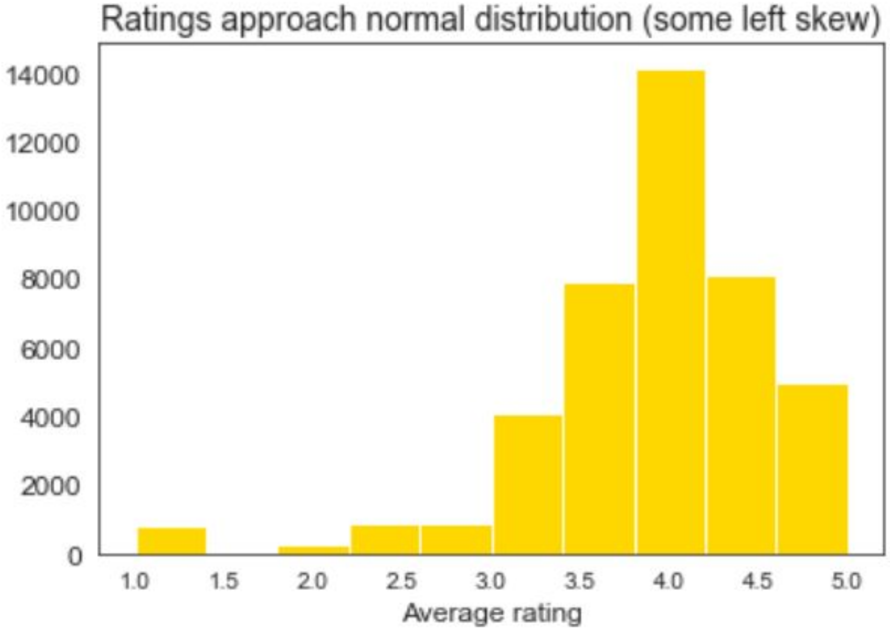
Abundance of top notes data

=

Project focus

☆ *Quick hit (top) only*

Data profile - Average ratings



After binning, modeled just on ratings 3 and 4

Image source: prada.com

EDA - Fragrance notes

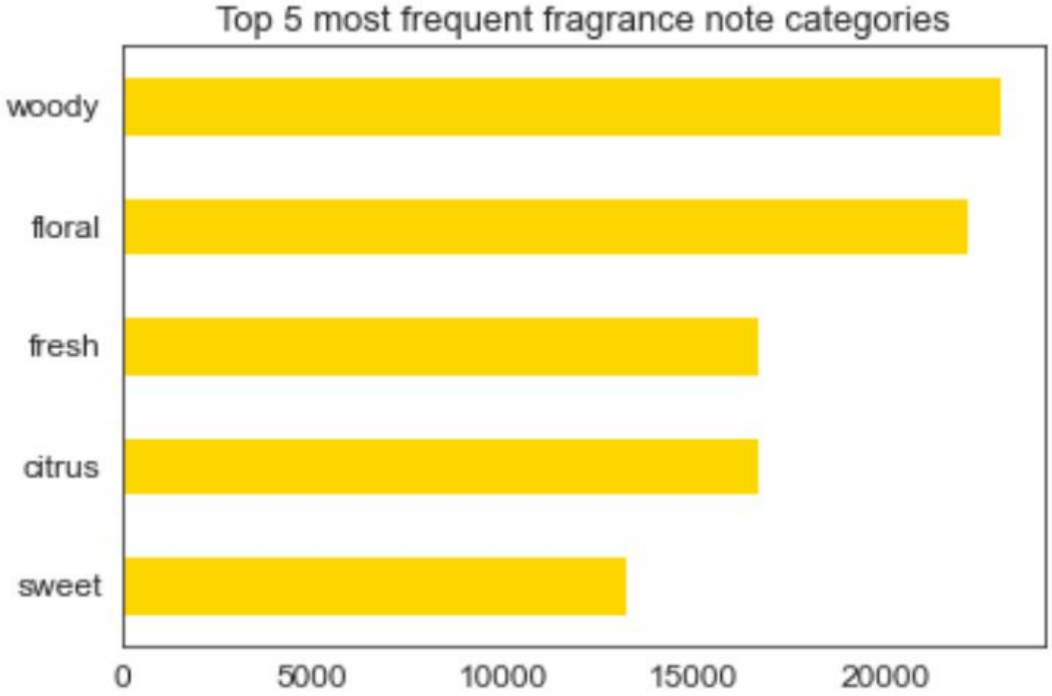
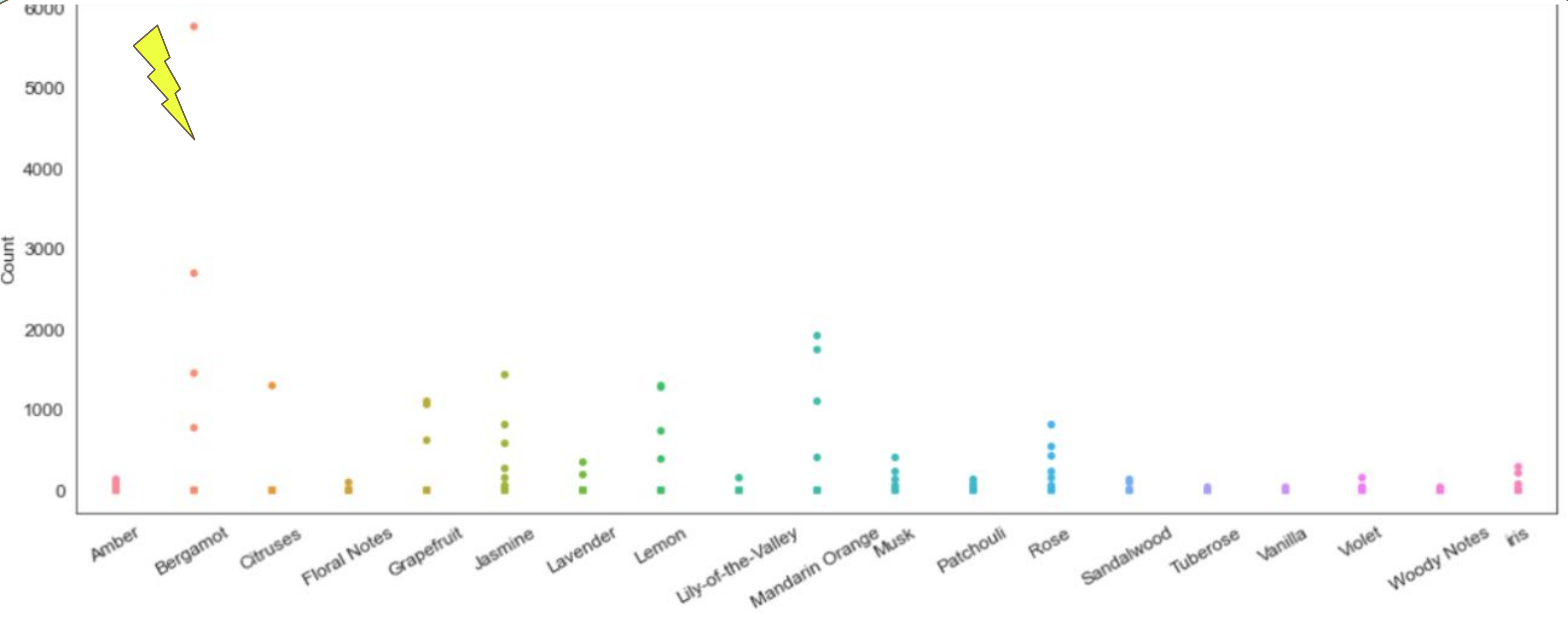


Image source: prada.com

EDA - Counts of most frequently used notes



Sampling of data science methods

A bit like sampling fragrances!

Logistic regression

Clustering

Image source: https://www.dior.com/fr_fr/parfum/maison-christian-dior/service-dessai-maison-christian-dior

Modeling results - Accuracy scores

Logistic regression
with *and* without
Principal Component Analysis (PCA)

Baseline = 54%

Majority class = Rating 4

With PCA	Without PCA
Train set: 59.6%	Train set: 91.4%
Test set: 56.9%	Test set: 52.3%
40 features	

Modeling results - Accuracy scores

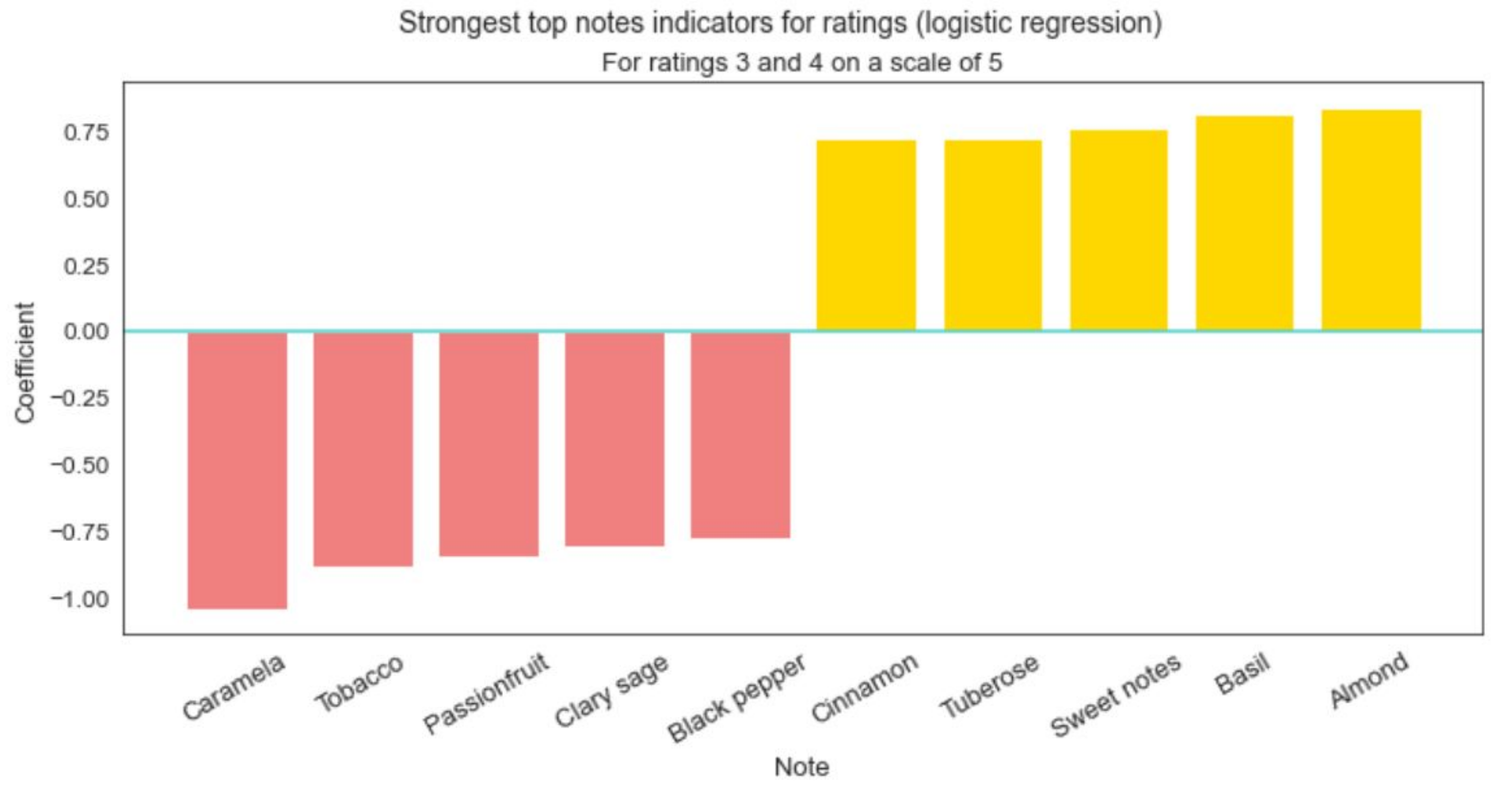
Tree-based models

Baseline model = 54%

Majority class = Rating 4

Decision tree	**Random Forest + GridSearchCV**
Train set: 92%	Train set: 92%
Test set: 54%	Test set: 60%

Modeling results: What we can infer



Clustering - Results (on fives ratings only)

Silhouette score

Cohesion *(intra-cluster distance)* - separation *(inter-cluster distance)*

Range: -1 (worst) to 1 (best)

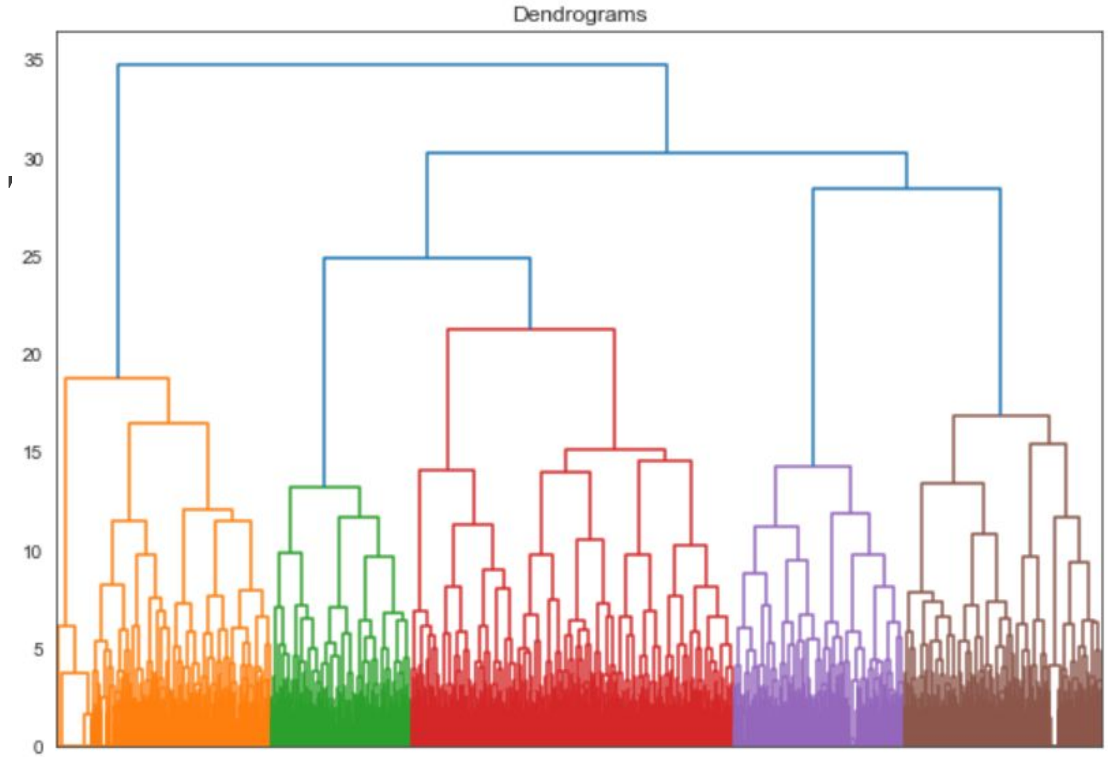
DBSCAN

Silhouette score: -0.3

*K*Means

Silhouette score: -0.2

Clustering - Next steps



Feature agglomeration

+

Hierarchical clustering

Conclusions and next steps (seriously)

Raw notes data has low predictive value

Tree-based system for product development

Live data stream

Now, a quick demo!

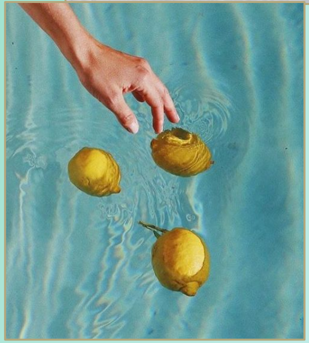


Image source: @fatimah.hsh and @taylor.g.simpson on @fruitassembly IG