

Decoding the Scent Space:

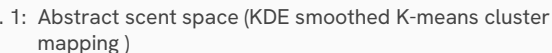
Revealing Hidden Structures in Fragrance Data

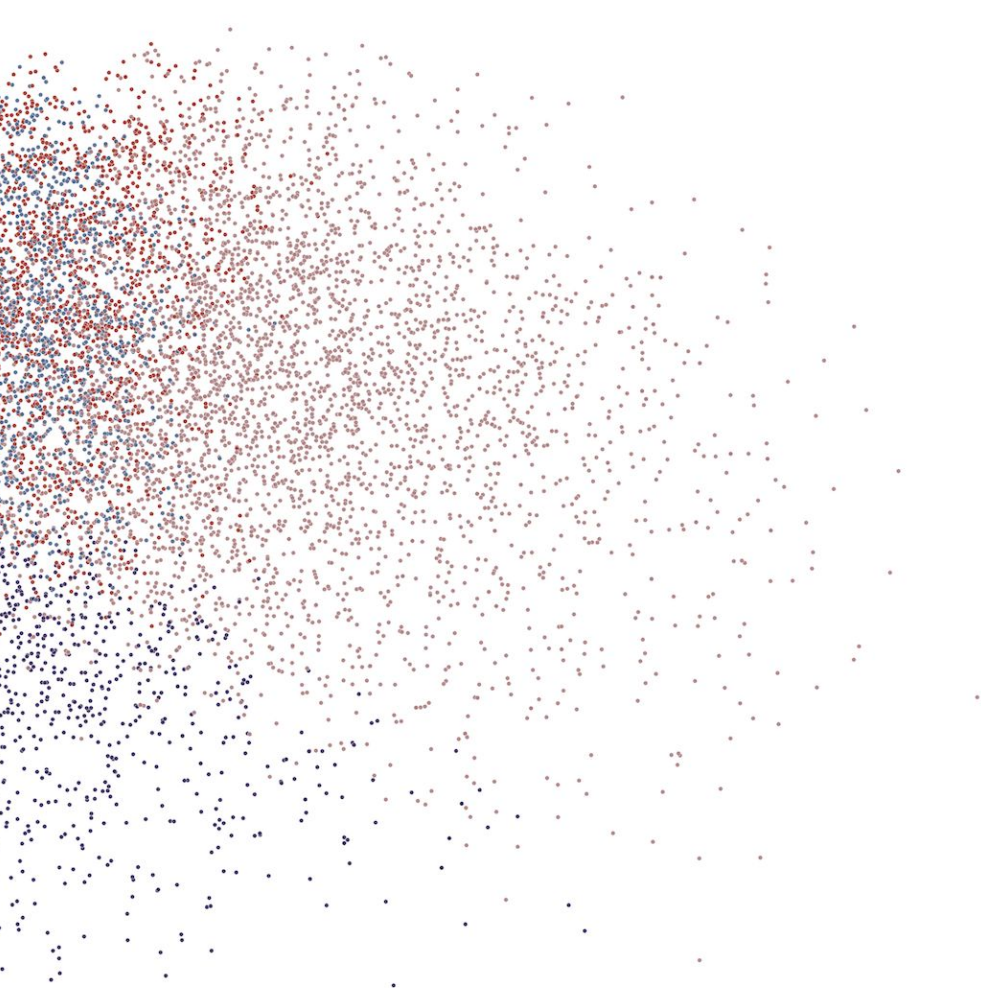
Benjamin Madden

December 2025

Fragrance is both chemistry and emotion.
A creative industry deserves a creative analytical approach.

A creative industry deserves a creative analytical approach.





Project Aim:

Map perfume formulations to reveal natural scent segments.

Identify ingredient trends and strategic opportunities for brands.

- Build a human-readable scent map from ingredient profiles
- Discover clusters that represent real olfactory 'neighbourhoods'
- Analyse ingredient trends over time
- Translate findings into insights for product, marketing, and innovation teams

Data Overview:

Source: doevent/perfume dataset (Hugging Face)
Raw dataset: 26,319 perfumes
Cleaned dataset: 26,248 perfumes (normalised, deduplicated, canonical notes)

Key fields used:

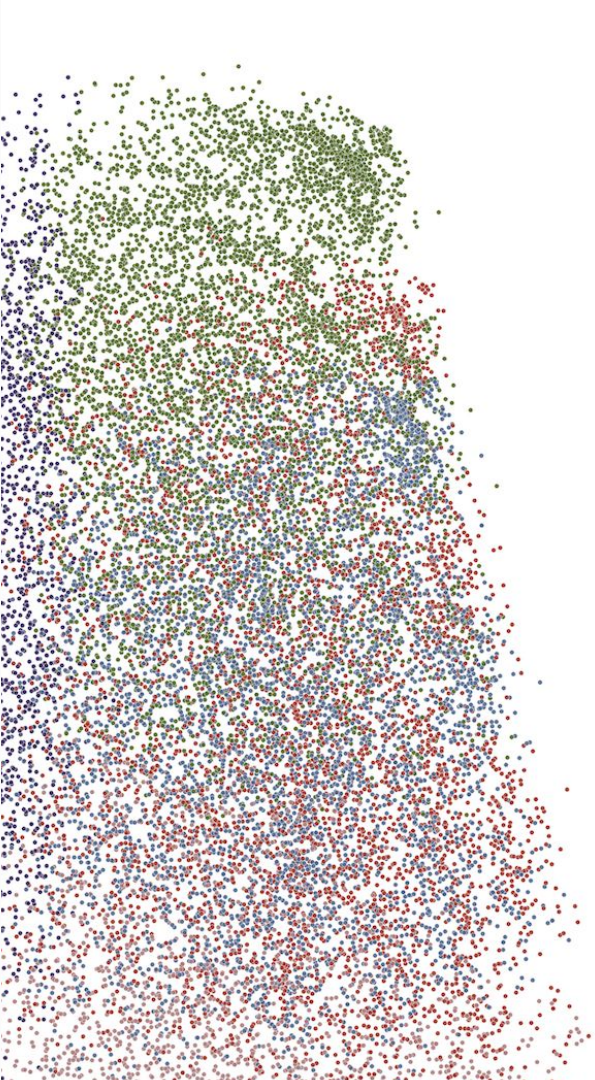
- ingredients_mapped
- perfume_id
- family
- years (1990-2025 majority)

Cleaning process:

- Standardised note formats
- Mapped synonyms and rare notes
- Removed inconsistent or incomplete rows.
- Removed fractured columns that were not necessary for clustering

Cleaned Dataset:

df.head(5)										
	brand	name_perfume	family	subfamily	fragrances	ingredients	gender	years	perfume_id	ingredients_mapped
0	fiorucci	wallstreet	floral	ambery (oriental)	floral amber fresher	[jasmine, lemon, spicy note, gardenia, rose, b...	male	2015	fiorucci - wallstreet	[jasmine, lemon, spicy note, gardenia, rose, b...
1	givenchy	désinvolte	floral	floral	floral fresher white flower	[orange blossom, jasmine, vetiver, magnolia, t...	unisex	2021	givenchy - désinvolte	[orange blossom, jasmine, vetiver, magnolia, t...
2	issey miyake	l'eau d'issey shade of sunrise 2019	floral	floral	floral fresher white flower	[jasmine, ylang ylang, heliotrope, osmanthus, ...	female	2019	issey miyake - l'eau d'issey shade of sunrise ...	[jasmine, ylang ylang, heliotrope, osmanthus, ...
3	guess	guess night	aromatic fougere	aromatic fougere	aromatic fougère fresher aromatic	[geranium, chilli, patchouli, ciste labdanum, ...	male	2013	guess - guess night	[geranium, chilli, patchouli, ciste labdanum, ...
4	tfk	signature line : n° 32	floral	citrus	floral crisp citrus fruity	[bergamot, iris, jasmine, neroli, rose, aldehy...	unisex	2015	tfk - signature line : n° 32	[bergamot, iris, jasmine, neroli, rose, aldehy...



Methods: From Formula to Scent Map

Pipeline:

- Ingredient normalisation → multi-hot vectors
- TruncatedSVD (50D) to capture structure
- PCA (2D) to create a human-readable map
- K-Means clustering (k = 5) to discover scent universes
- Trend detection across notes and families over time

How K-Means clustering works

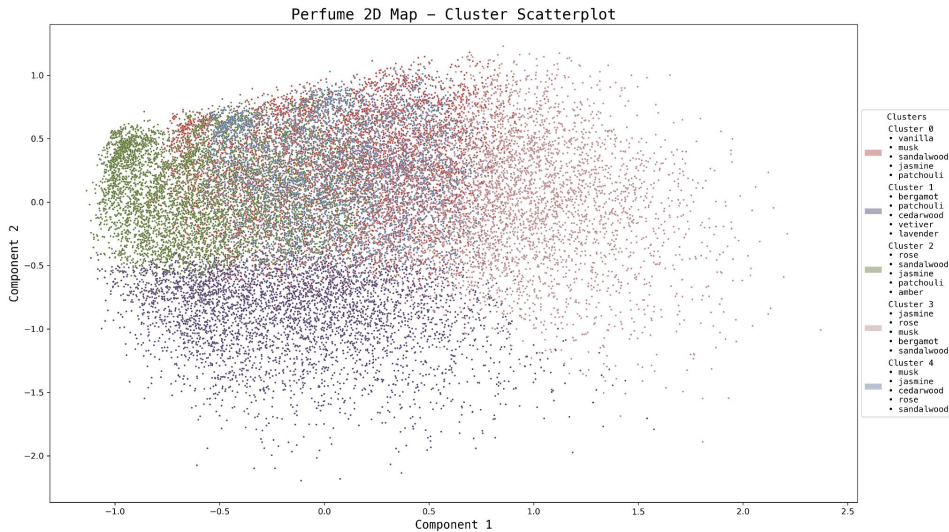
- Groups perfumes by ingredient similarity
- Each cluster has a centre point – the ‘average’ scent profile
- Perfumes are assigned to the nearest centre point
- Centre points adjust iteratively until clusters become stable

Reproducibility:

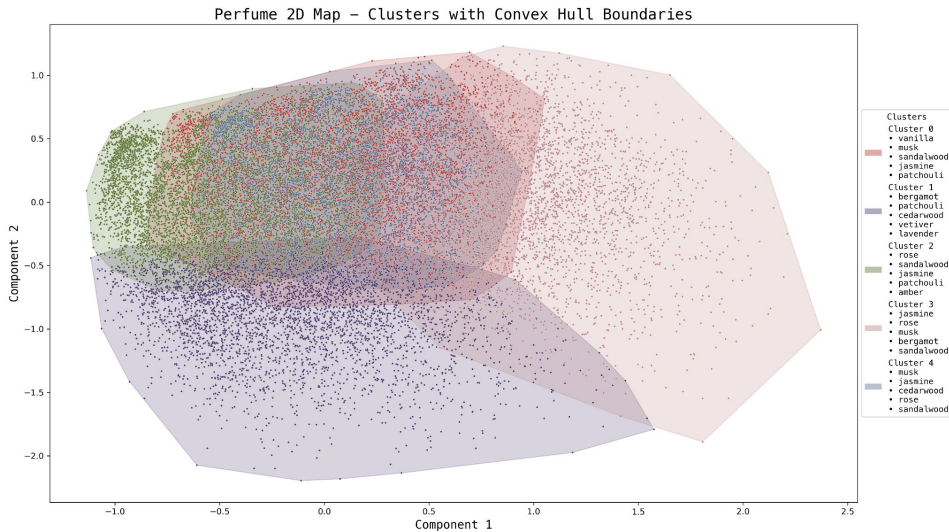
All steps were stored as part of a reproducible pipeline.

	k	silhouette	davies_bouldin	inertia
0	2	0.060135	4.042913	123385.827342
1	3	0.047619	3.901831	119231.280650
2	4	0.031390	3.658780	116239.246086
3	5	0.026260	3.514880	114010.687582
4	6	0.022905	3.672429	112207.770453
5	7	0.025295	3.639720	110815.871907
6	8	0.017800	3.530891	109511.184988
7	9	0.018969	3.392854	108445.051485
8	10	0.016194	3.421730	107441.667498
9	11	0.013847	3.383367	106341.661832
10	12	0.016551	3.424041	105674.248858

The 'Shape of Scent'



Scatterplot of perfumes in 2D ingredient space. Points close together share similar ingredient profiles; distant points are more distinct.



Scatterplot with convex hulls highlighting cluster boundaries. Provides clear visualisation of olfactory segments.

Metrics & Limitations

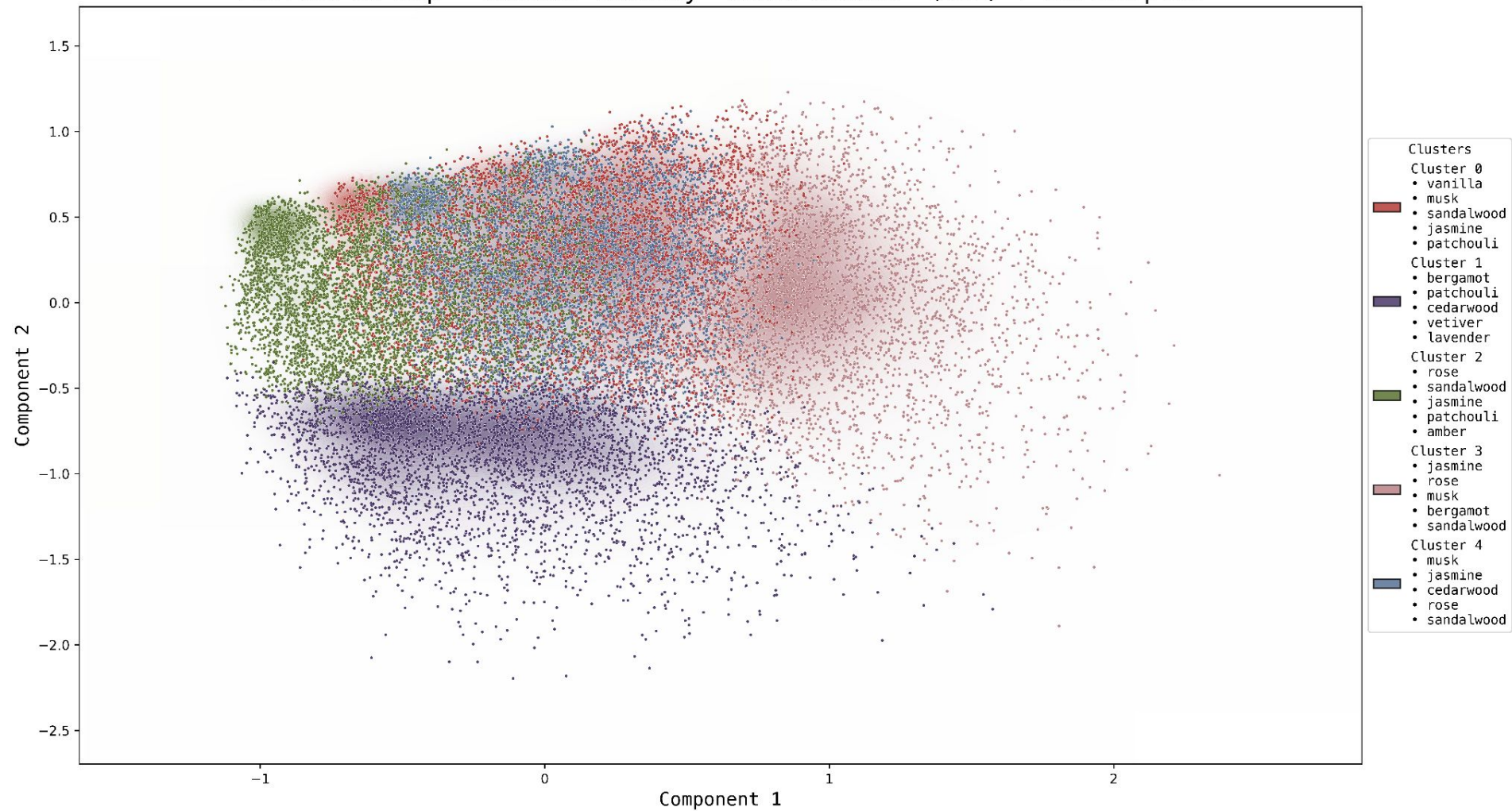
Model metrics:

- Silhouette (sample cosine): 0.017
- Davies-Bouldin: ~3.5
- SVD 50-components capture structural variance; 2D is illustrative

Limitations:

- Ingredients vary in completeness
- Fragrance families are marketing-driven
- 2D compression simplifies a complex space
- No sales or consumer sentiment data

Perfume 2D Map – Cluster Density ‘Scent Clouds’ (KDE) + Scatterplot



Cluster Highlights: Scent Clouds in Profile

Cluster 0 — Warm Gourmand Woods

Top notes: vanilla, musk, sandalwood, amber
Character: cosy, sweet, warm, enveloping
Market: largest cluster; very saturated
Opportunity: differentiate through texture
(smoke, resin, mineral)

Cluster 1 — Fresh Citrus-Aromatic Woods

Top notes: bergamot, patchouli, cedarwood,
vetiver, lavender
Character: crisp, herbal, energetic
Market: stable masculine territory
Opportunity: modernised fougères, lighter
unisex blends

Cluster 2 — Green Floral-Woody

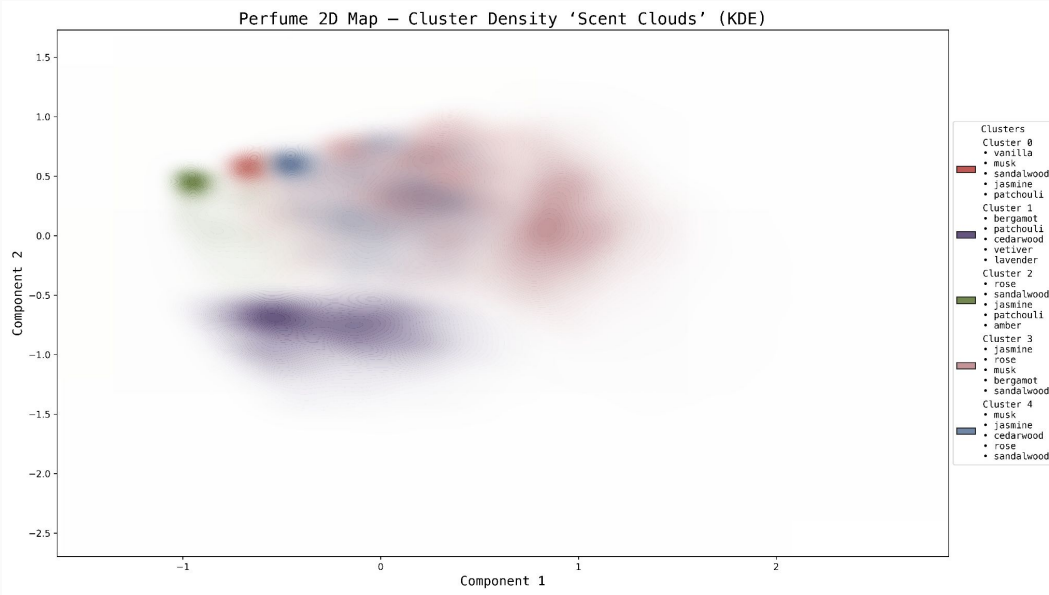
Top notes: rose, sandalwood, jasmine,
patchouli, amber
Character: natural, leafy, balanced
Market: strong unisex crossover
Opportunity: artisanal florals, botanical
woods, niche storytelling

Cluster 3 — Bright Florals with Musk

Top notes: jasmine, rose, musk,
sandalwood
Character: soft, clean, contemporary
Market: extremely crowded
Opportunity: transparency, rare florals,
fresh interpretations

Cluster 4 — Clean Woody Musks

Top notes: musk, jasmine, cedarwood,
rose, sandalwood
Character: airy, minimalist, skin-like
Market: moderate density but rising
Opportunity: expand into “second-skin”
and mineral woody profiles



Emerging Ingredient Trends

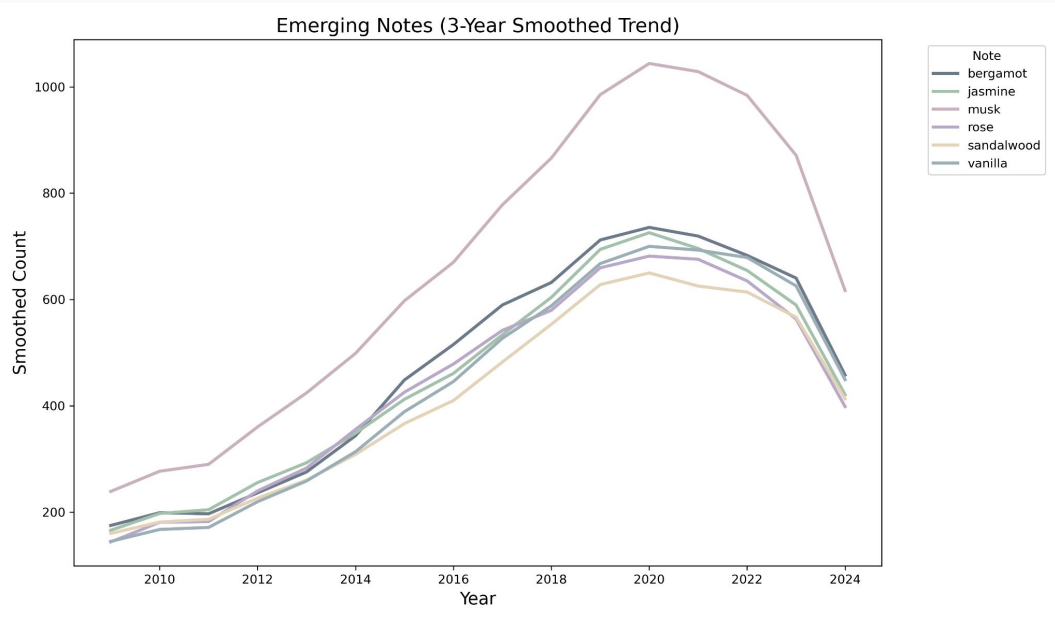
- Identified via year-over-year patterns
- The last 15 years show the strongest movement
- Key emerging differentiator notes:
 - Cardamom*
 - Fig*
 - Tea*
 - Pink pepper*
 - Ambergris*
- Align with growth in Clusters 2 and 4
- The visual shows the top emerging notes with enough yearly data for clear trend lines

note	first_sum	last_sum	ratio_last_first	pct_change_last_first	avg_yoy	years_observed	category
aldehyde	5	111	22.2	21.2	0.06484129126678170	104	emerging
amber	1	2332	2332.0	2331.0	-0.026425929599859200	197	emerging
ambergris	1	825	825.0	824.0	0.031891896263216900	150	emerging
ambrette seed	1	183	183.0	182.0	-0.0014466809412387100	119	emerging
angelica	1	98	98.0	97.0	0.0263680667382229	91	emerging
animalic note	1	44	44.0	43.0	0.004778807051534330	99	emerging
anise	1	81	81.0	80.0	-0.024407090085056200	118	emerging
apple	4	406	101.5	100.5	0.1953623164720150	34	emerging
apricot	10	115	11.5	10.5	0.2547077342861520	37	emerging
aromatic note	1	67	67.0	66.0	-0.02996601786346460	114	emerging
artemisia	1	100	100.0	99.0	-0.01862166048529130	70	emerging
balsam	1	55	55.0	54.0	-0.00482120646348774	197	emerging
bamboo	10	38	3.8	2.8	0.37459743321812300	29	growing

[Extract from notes categorization data]

Categorises into:

Emerging
Growing
Stable



Strategy & Next Steps

Recommendations

1. Fast follow launches:
Target emerging notes within growing clusters (2 and 4)
2. White space exploration:
Leverage sparse outer regions for niche or artisan concepts
3. Brand strategy alignment:
Match cluster personas to brand DNA to guide future releases

Next steps

- Consumer validation & scent lexicon development
- Incorporate review sentiment + fine-grained release data
- Extend to brand-level competitive mapping