Predicting Ice cream Deliciousness

Composition and Online Ratings of Vanilla Ice Cream

A Cross-Sectional Study of 5,793 Reviews and 37 USDA-Matched Products

Bob Zhao
The Webb Schools

Abstract

Vanilla ice cream looks simple, but a lot is happening under the lid. This paper asks whether the parts shoppers can easily see on a label—especially fat, sugar, and ingredients—help predict how much people like a product. I linked a public archive of manufacturer-site reviews to the October 2020 USDA FoodData Central release. The strict sample has 45 spoonable, vanilla-labeled products from four brands and 5,793 star ratings. Thirty-seven products could be matched conservatively to nutrition records. To avoid giving tiny review samples too much influence, I used a product-level empirical-Bayes rating. The main regression tested linear and curved relationships for fat and total sugars, while controlling for brand. The four composition terms were not jointly related to ratings ($F_{4,29} = 0.97$, $p = 0.438$), adjusted R^2 was -0.066, and every confidence interval crossed zero. In leave-one-product-out testing, both the linear and quadratic models predicted worse (RMSE 0.411 and 0.428) than simply using the training-sample mean (0.378). Ingredient count and seven ingredient flags also produced no stable findings after multiple-test correction or in a broader 130-product check. Review wording told a clearer story. Within the same product, creamy or smooth wording was associated with 0.383 more stars, while icy or hard wording was associated with 1.414 fewer stars. The answer is not that ingredients do not matter. It is that a nutrition label is a pretty bad crystal ball for ice cream enjoyment when processing, air, storage, mix-ins, branding, expectations, and reviewer selection are missing.

Keywords: vanilla ice cream; online reviews; nutrition labels; FoodData Central; regression; texture

1 Background

1.1 Vanilla is not actually boring

Vanilla is often treated as the default button of ice cream. That makes it a useful test case: the base dairy mixture is easier to notice than it is in a flavor like double chocolate brownie. Still, a pint of vanilla is doing a surprising amount of physics while sitting quietly in the freezer. It contains ice crystals, air bubbles, fat droplets, dissolved sugar, milk proteins, flavor compounds, stabilizers, and emulsifiers. Small changes in freezing, air content, storage, and fat structure can change hardness, melting, creaminess, and flavor release. Studies of commercial vanilla ice creams have found that these sensory properties help drive consumer acceptance [3].

The category is messier than its name suggests. A product labeled with vanilla might be plain vanilla, French vanilla, vanilla bean, gelato, reduced-sugar dessert, non-dairy dessert, or vanilla plus cookies, caramel, nuts, and other very non-vanilla guests. In this paper, "vanilla ice cream" is shorthand for spoonable, vanilla-labeled frozen desserts. Not every item necessarily meets the legal U.S. definition of ice cream. That standard generally requires at least 10 percent milkfat and 10 percent nonfat milk solids, with specific adjustments, plus a minimum finished weight of 4.5 pounds per gallon [7]. I kept the broader shopper-facing category because that is what people actually see online.

1.2 What fat and sugar might do

Fat can add body, lubrication, flavor delivery, and protection from iciness. Sugar adds sweetness, but it also lowers the freezing point. This means that "more is better" is not an especially convincing theory for either nutrient. In a controlled experiment with nine vanilla formulas and 146 consumers, Guinard et al. [1] found curved liking patterns over products containing 8.73–19.30 percent fat and 8.94–18.81 percent sugar. Their predicted favorite was near 14.77 percent fat and 14.30 percent sugar. That study is why I tested squared terms instead of forcing the relationship to be a straight line.

Controlled formulas and store products answer different questions, though. Rolon et al. [4] found similar acceptability across vanilla ice creams with 6–14 percent fat when maltodextrin replaced some of the removed fat. A company can change fat and also change sweeteners, stabilizers, air, processing, or flavor. The number printed on the nutrition label is therefore part of a recipe bundle, not a clean experiment.

1.3 Online reviews: lots of data, some baggage

Online reviews are useful because they record reactions from thousands of people instead of a small tasting panel. They are also self-selected. People decide whether to review, and very happy or very unhappy customers may be more motivated than everyone in the quiet middle. Research on online product reviews has documented this selection problem [2]. In the full source archive here, 71.0 percent of the 21,674 reviews are five stars. In other words, these reviewers really liked the five-star button.

There is another important catch. Fat and ingredients change between products, not between two reviewers of the same product. Calling all 5,793 reviews independent fat experiments would make the sample look much larger than it really is. I therefore used products as the units in the main analysis. Individual reviews only appear in a secondary analysis of the words reviewers used.

1.4 Research questions

I asked three questions:

1. Do fat and total sugars have the curved, inverted-U relationship with ratings suggested by controlled sensory research?
2. Do ingredient-list length or declared features such as gums, emulsifiers, corn syrup, egg, vanilla bean, vanilla extract, and alternative sweeteners explain ratings?
3. Within the same product, are words about creaminess, hardness, artificial taste, sweetness, texture, vanilla, or mix-ins connected to star ratings?

The first question is the main statistical test. The ingredient tests are exploratory and include a multiple-testing correction. The word analysis is a second viewpoint, not proof that a texture word caused a rating. Because all of the data are observational, I describe relationships rather than causes.

2 Data

2.1 Reviews and products

The review data come from version 3 of the public *Ice Cream Dataset* on Kaggle [5]. The final version was updated on October 4, 2020 and is marked CC0 1.0. It contains 241 products and 21,674 reviews collected from the product sites of Ben & Jerry's, Häagen-Dazs, Breyers, and Talenti. The files include product names, descriptions, ingredients, displayed ratings, individual star ratings, dates, review text, and helpfulness counts.

I first checked whether the product and review tables agreed. The downloaded review count exactly matched the displayed count for 228 of 241 products. Across the remaining products, the total absolute difference was only 20 reviews. Recalculated means differed from the displayed ratings by 0.0247 stars on average and never by more than 0.05. These checks cannot prove that every reviewer was genuine, but they show that the files fit together properly.

2.2 Building the sample

A product entered the vanilla-labeled pool if *vanilla* appeared in its name or subhead. This produced 51 products and 6,190 reviews. I excluded six bars, squares, and snack cups because their coating, shape, serving style, and surface area make them awkward comparisons with a tub of ice cream. The strict sample has 45 spoonable products and 5,793 reviews dated November 20, 2015 through October 2, 2020. Thirteen products are plain vanilla by a reproducible name rule, while the other 32 include another flavor or mix-in. Figure 1 shows the full filtering process.

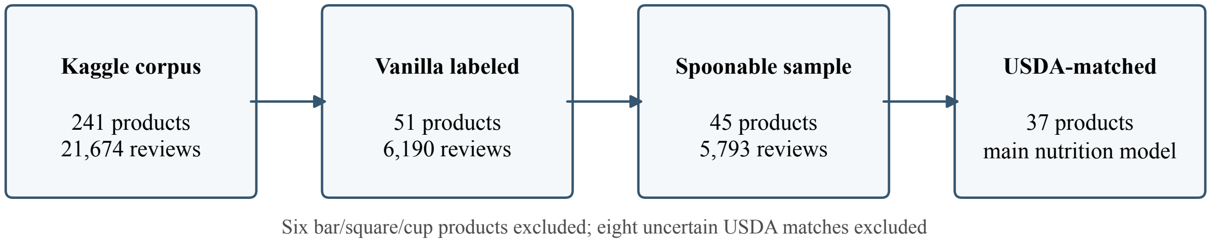

Figure 1: How the analytical samples were built. The first three stages use the Kaggle archive. The last stage requires a conservative match to the October 2020 USDA nutrition snapshot.

Table 1: Strict analytical sample by brand

Brand	Products	Reviews	Median reviews	Mean EB rating	USDA matches
Ben & Jerry's	13	2,834	72	4.27	13
Breyers	16	1,630	80	4.25	11
Haagen-Dazs	9	787	74	4.20	7
Talenti	7	542	65	4.20	6
Total	**45**	**5,793**	72	4.24	**37**

Notes: The empirical-Bayes (EB) rating uses a prior strength of 20 reviews. "USDA matches" are the products used in nutrition regressions. Review counts are downloaded rows.

Each brand contributes 7–16 products (Table 1). Product review counts are uneven: the median is 72, the middle 50 percent runs from 38 to 115, and the full range is 10–978. Ben & Jerry's contributes almost half of the reviews. I included brand indicators in the main models but still gave each product one vote.

For a broader ingredient check, I also selected every product in the full 241-product archive whose ingredient list contains *vanilla*. That produces 130 products. Vanilla may be a minor ingredient in some of them, so this broader group is only a replication test, not the primary sample.

2.3 Nutrition data and matching

Nutrition information comes from the October 30, 2020 USDA FoodData Central Branded Foods release [6]. This date is close to the end of the review archive, which is much better than attaching today's label to a review written years ago. The USDA files report label values per 100 grams. I used total fat, total sugars, protein, saturated fat, sodium, and energy when available. Missing values stayed missing; they were never quietly changed to zero.

The review file and USDA file do not share a UPC or other product ID, so the matching required some caution. I restricted candidates to verified names for the same brand owners. I then compared normalized ingredient words and product-name words. If J_I is ingredient similarity and J_N is name similarity, the ranking score was

$$M = 0.80J_I + 0.20J_N.$$

The top match was accepted automatically only under one of three rules: $J_I \geq 0.98$ and $J_N \geq 0.15$; $J_I \geq 0.85$ and $J_N \geq 0.20$; or $J_I \geq 0.65$ and $J_N \geq 0.40$. These rules kept a shared base recipe from being attached to the wrong flavor. Every accepted match, rejected match, runner-up, and score is saved in the audit CSV. Two borderline pairs—Vanilla Caramel Gelato and Coffee Vanilla Chocolate Trio—were accepted after side-by-side checks of their names, brands, and formulas. No other manual override was used.

The final nutrition sample has 37 products: 24 near-exact ingredient matches and 13 other high-confidence matches. Eight uncertain pairs were left out. The accepted and excluded products had similar average EB ratings (4.249 versus 4.187; Welch $p = 0.668$), although eight exclusions are too few to prove that missing matches are harmless.

2.4 What happened to price?

Price was tempting. It was also a data trap. The public review archive has no exact-SKU, store-level 2020 prices. Adding a 2026 grocery price to reviews from 2015–2020 would mix inflation, sales, package-size changes, location, and possible recipe changes. That number would look precise while being wrong in several directions at once. I left price out and describe a better price design later in Section 5.5.

3 Methods

3.1 A fairer product rating

Some products have nearly 1,000 reviews and others have only 10. I did not want one tiny review sample to become the surprise champion because of a lucky streak. Let n_{is} be the number of reviews that give product i a star value $s \in \{1, \dots, 5\}$, and let q_s be that star's proportion across all 21,674 reviews. With a prior strength of $m = 20$, I calculated

$$y_i = \sum_{s=1}^{5} s \left(\frac{n_{is} + mq_s}{n_i + m} \right).$$

This empirical-Bayes (EB) score pulls small samples a little toward the platform mean of 4.224 stars. Products with hundreds of reviews barely move. I also repeated the analysis with $m = 5$, $m = 50$, and raw product averages. The regression is unweighted because 500 more reviews do not create another ice cream formula.

3.2 Composition and ingredient measures

The main composition variables are total fat and total sugars in grams per 100 grams. I centered each variable and divided it by its sample standard deviation before adding squared terms. Fat and sugar have a moderate correlation in the matched sample ($r = 0.473$), so their coefficients need to be read together.

Ingredient count is the number of top-level comma- or semicolon-separated entries, while ignoring punctuation inside parentheses. It is a consistent counting rule, not a chemistry score. Binary flags record gums or carrageenan, emulsifiers, corn syrup, egg or egg yolk, vanilla bean or seed, vanilla extract, and alternative sweeteners. High-fructose corn syrup appears in only one strict-sample item and explicit artificial flavor appears in none, so I did not test those two flags.

3.3 Main regression

The primary model was

$$y_i = \beta_0 + \beta_1 F_i + \beta_2 F_i^2 + \beta_3 S_i + \beta_4 S_i^2 + \sum_{b=1}^{3} \alpha_b Brand_{ib} + \varepsilon_i,$$

where F_i and S_i are standardized fat and sugar. Three brand indicators account for average differences among the four brands. A true inverted-U pattern would usually make the squared terms negative. The main decision comes from a joint test of $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, not from hunting for one exciting-looking coefficient.

I used ordinary least squares with HC3 standard errors, which are more cautious when the amount of variation is uneven across observations. Confidence intervals and p values use the remaining degrees of freedom. The model is descriptive: it does not imitate a factory experiment where only one nutrient changes and everything else stays fixed.

3.4 Ingredient tests, prediction, and review words

The compact ingredient model includes ingredient count, plain-vanilla status, reduced/modified status, and brand. Then I tested each of seven ingredient flags separately with plain-vanilla and brand controls. I adjusted the eight exploratory p values with the Benjamini-Hochberg false-discovery procedure at 0.05. The same models were repeated in the broader 130-product pool.

To test whether the models work beyond the data used to fit them, I performed leave-one-product-out cross-validation. For each of the 37 matched products, I refit the model on the other 36 and predicted the missing one. I compared a training-mean baseline, a linear fat-and-sugar model, and the main quadratic model using RMSE, MAE, and predictive R^2. I also left out one whole brand at a time as a tougher stress test.

Finally, review titles and bodies were searched with seven fixed dictionaries: creamy/smooth; too sweet; artificial/fake/chemical; icy/hard/freezer-burn; texture or consistency; vanilla/vanillin; and mix-ins. For each theme k, I fit

$$Stars_{ij} = \mu_i + \gamma_k Theme_{ijk} + u_{ij},$$

where μ_i is a product fixed effect. In plain English, this compares two reviews of the same product when one mentions the theme and the other does not. Standard errors are clustered by product, and the seven theme tests are adjusted for multiple comparisons. The dictionaries miss sarcasm, negation, and context, so this is a useful clue rather than mind reading.

3.5 Checks and reproducibility

Sensitivity tests changed the EB prior, used raw ratings, added plain-vanilla and reformulation indicators, removed brand indicators, and kept only the 24 near-exact nutrition matches. I also calculated rank correlations. The supplied Python scripts create every selection, model, correction, table, and figure. Archive hashes make it possible to check that the same source files were used.

4 Results

4.1 What the sample looks like

The 45 products have an average EB rating of 4.24 stars, a standard deviation of 0.37, and a range of 3.06–4.73. Among the 37 matched products, fat averages 12.83 g/100 g (range 5.17–20.33) and total sugars average 21.67 g/100 g (range 5.13–29.06). The products cover a useful range, but companies chose the whole formula at once. This is still a freezer aisle, not a randomized laboratory.

Table 2: Descriptive statistics

Variable	N	Mean	SD	Median	Min.	Max.
EB-adjusted rating	45	4.24	0.37	4.32	3.06	4.73
Raw mean rating	45	4.26	0.48	4.35	2.96	4.97
Reviews per product	45	128.73	195.57	72.00	10.00	978.00
Top-level ingredient count	45	17.27	7.27	17.00	5.00	38.00
Fat (g/100 g)	37	12.83	4.28	13.48	5.17	20.33
Total sugars (g/100 g)	37	21.67	5.25	23.26	5.13	29.06
Protein (g/100 g)	37	3.91	0.88	3.81	2.47	6.76
Saturated fat (g/100 g)	37	7.78	2.43	7.75	3.38	13.82
Sodium (mg/100 g)	37	74.03	27.42	67.00	45.00	149.00

Notes: Ratings, review counts, and ingredient count use all 45 strict-sample products. Nutrients use the 37 accepted USDA matches and are reported per 100 g.

Figure 2 shows the unadjusted product patterns. Bubble size grows with the logarithm of review count. The dashed lines are only descriptive curves, not the adjusted main model. Brand clusters are visible, and the low-sugar end contains only a few modified products. That sparse corner deserves caution, not a dramatic trend line.

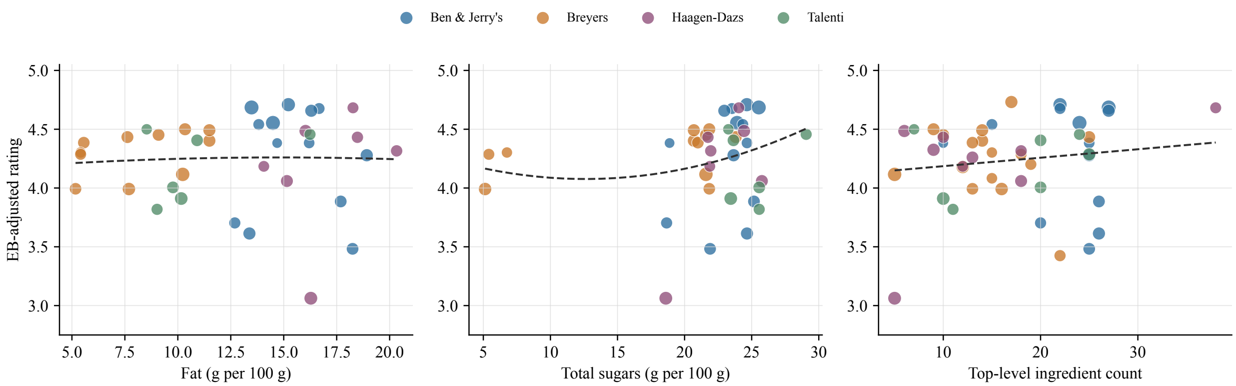

Figure 2: Composition and EB-adjusted ratings. Fat and sugars use the 37 accepted matches; ingredient count uses all 45 products. Point size shows log review count. Dashed curves are unadjusted summaries.

4.2 Fat and sugar do not predict ratings well

Table 3 gives the main quadratic regression. A one-standard-deviation increase in fat at the average sugar value is associated with only 0.055 rating points (95% CI −0.225 to 0.336; $p = 0.689$). The fat-squared estimate is nearly flat at −0.004 (95% CI −0.163 to 0.154; $p = 0.956$). The sugar estimate is larger but uncertain: 0.274 points per standard deviation (95% CI −0.174 to 0.722; $p = 0.221$). The sugar-squared estimate is positive, not the predicted negative direction, at 0.072 (95% CI −0.069 to 0.213; $p = 0.304$).

Table 3: Primary quadratic model of EB-adjusted product rating

Predictor	Estimate	HC3 SE	95% CI	p
Fat (1 SD)	0.055	0.137	[-0.225, 0.336]	0.689
Fat^2	-0.004	0.077	[-0.163, 0.154]	0.956
Total sugars (1 SD)	0.274	0.219	[-0.174, 0.722]	0.221
Total sugars^2	0.072	0.069	[-0.069, 0.213]	0.304

Notes: $N = 37$. Fat and sugars are centered and scaled by 4.279 and 5.252 g/100 g, respectively, before squaring. The model includes brand indicators. HC3 intervals use 29 residual degrees of freedom. The four composition terms are jointly nonsignificant: $F(4, 29) = 0.971$, $p = 0.438$.

The four composition terms fail the joint test ($F_{4,29} = 0.971$, $p = 0.438$). The model has $R^2 = 0.141$, but adjusted $R^2 = -0.066$. In other words, once the model pays a penalty for all its extra terms, it has not earned its keep. Figure 3 shows that every 95 percent confidence interval crosses zero. This does not prove the effects are exactly zero; it says this sample cannot support a dependable label-based curve.

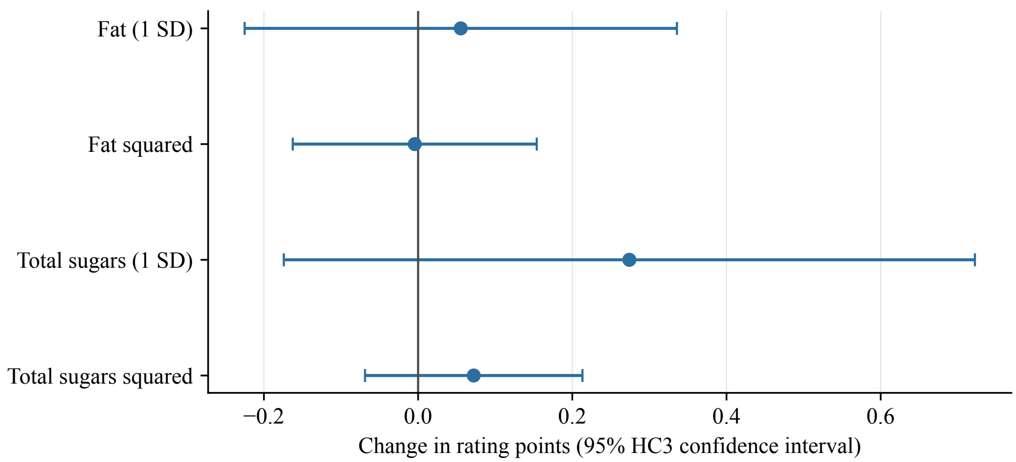

Figure 3: Main composition coefficients with 95% HC3 confidence intervals. Every interval crosses zero.

The model-light rank correlations agree. Fat and EB rating have Spearman $\rho = 0.150$ ($p = 0.376$), while sugars and rating have $\rho = 0.189$ ($p = 0.261$).

4.3 Ingredients do not produce a secret formula

The average product lists 17.27 top-level ingredients, with a range of 5–38. In the compact model, a one-standard-deviation increase in ingredient count corresponds to 0.038 rating points (95% CI -0.178 to 0.253; $p = 0.726$). Plain vanilla is associated with -0.069 points ($p = 0.719$), while reduced/modified positioning is associated with 0.014 points ($p = 0.940$). Adjusted R^2 is -0.125.

Egg or egg yolk has a nominal positive result in the strict sample ($b = 0.251$, unadjusted $p = 0.025$), but it does not survive correction across eight tests (BH $q = 0.198$). Its sign also flips in the broader sample ($b = -0.068$). No strict-sample ingredient feature reaches $q < 0.05$, and every broad-sample adjusted value is 0.992. So, sadly, there is no evidence for a simple “add this ingredient, gain stars” cheat code.

Table 4: Exploratory ingredient-feature models

Feature	N present	Strict b	p	BH q	Broad b	Broad q
Ingredient count (1 SD)	45	0.038	0.709	0.843	0.023	0.992
Gum or carrageenan	22	-0.073	0.852	0.852	0.126	0.992
Emulsifier	30	0.060	0.613	0.843	-0.001	0.992
Corn syrup	18	0.041	0.724	0.843	0.072	0.992
Egg or egg yolk	26	0.251	0.025	0.198	-0.068	0.992
Vanilla bean or seed	9	-0.119	0.520	0.843	-0.027	0.992
Vanilla extract	27	-0.089	0.658	0.843	0.263	0.992
Alternative sweetener	4	-0.044	0.738	0.843	-0.163	0.992

Notes: Each row is a separate HC3 regression of EB rating on the named feature, plain-vanilla status, and brand indicators. Strict $N = 45$; broad replication $N = 130$. BH q values adjust eight tests in each sample.

4.4 The prediction test is even less impressed

The training-mean baseline has a leave-one-product-out RMSE of 0.378. Adding linear fat and sugar makes RMSE worse at 0.411, and adding the squared terms makes it 0.428 (Table 5). Negative predictive R^2 values tell the same story. A flexible curve can look interesting in the data it already knows, but the real test is meeting a product it has never seen.

Table 5: Leave-one-product-out predictive performance

Model	LOO RMSE	LOO MAE	Predictive R^2
Training-mean baseline	0.378	0.301	-0.056
Linear composition	0.411	0.324	-0.251
Quadratic composition	0.428	0.352	-0.357

Notes: Metrics use the 37 matched products. Each composition model includes brand and is refit in every fold. Lower RMSE and MAE are better.

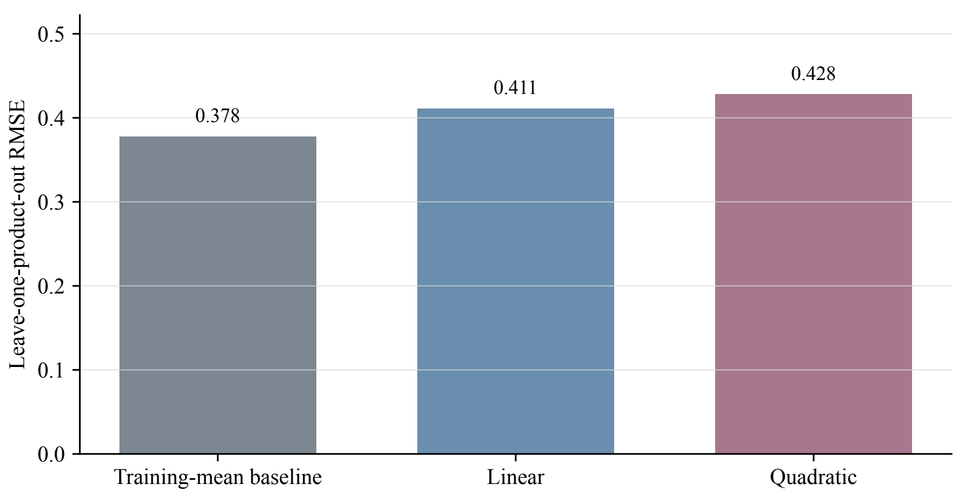

Figure 4: Leave-one-product-out RMSE. Both composition models lose to the training-mean baseline.

Leaving out an entire brand is harsher. RMSE is 0.370 for the training mean, 0.566 for the linear composition model, and 1.840 for the quadratic model. The last model has to extrapolate beyond the remaining brands' composition range and goes badly off course. That is a warning against announcing one "perfect" fat-and-sugar recipe from this sample.

4.5 Reviewers care about the spoon experience

The review-language analysis uses all 5,793 reviews while comparing language within the same product. Creamy or smooth wording appears in 18.4 percent of reviews and is associated with 0.383 more stars (95% CI 0.247 to 0.519; BH $q < 0.001$). Icy or hard wording appears in only 1.6 percent, but it has the biggest contrast: -1.414 stars (95% CI -1.824 to -1.005; $q < 0.001$).

Artificial or fake language is associated with -0.339 stars ($q = 0.017$), general texture language with -0.466 ($q < 0.001$), and mix-in language with -0.205 ($q < 0.001$).

Table 6: Within-product relationships between review themes and stars

Theme	Mentions	Reviews	b	95% CI	BH q
Creamy or smooth	1066	18.4%	0.383	[0.247, 0.519]	< 0.001
Too sweet	130	2.2%	-0.198	[-0.401, 0.006]	0.066
Artificial or fake	130	2.2%	-0.339	[-0.601, -0.078]	0.017
Icy or hard	90	1.6%	-1.414	[-1.824, -1.005]	< 0.001
Texture or consistency	501	8.6%	-0.466	[-0.637, -0.296]	< 0.001
Vanilla or vanillin	1891	32.6%	-0.168	[-0.360, 0.023]	0.084
Mix-ins	1677	28.9%	-0.205	[-0.291, -0.118]	< 0.001

Notes: Each row is a separate review-level model with product fixed effects and product-clustered standard errors. "Reviews" is theme prevalence among 5,793 reviews. BH q values adjust seven tests. A keyword can appear in praise or a complaint.

Too-sweet language is negative at -0.198 stars, but its interval barely crosses zero and it misses the corrected cutoff ($q = 0.066$). Vanilla/vanillin language is also unclear (-0.168, $q = 0.084$). Mix-in wording does not mean that mix-in products are bad overall. Within a specific product, people may bring up a mix-in because there was too little, too much, or the texture was disappointing.

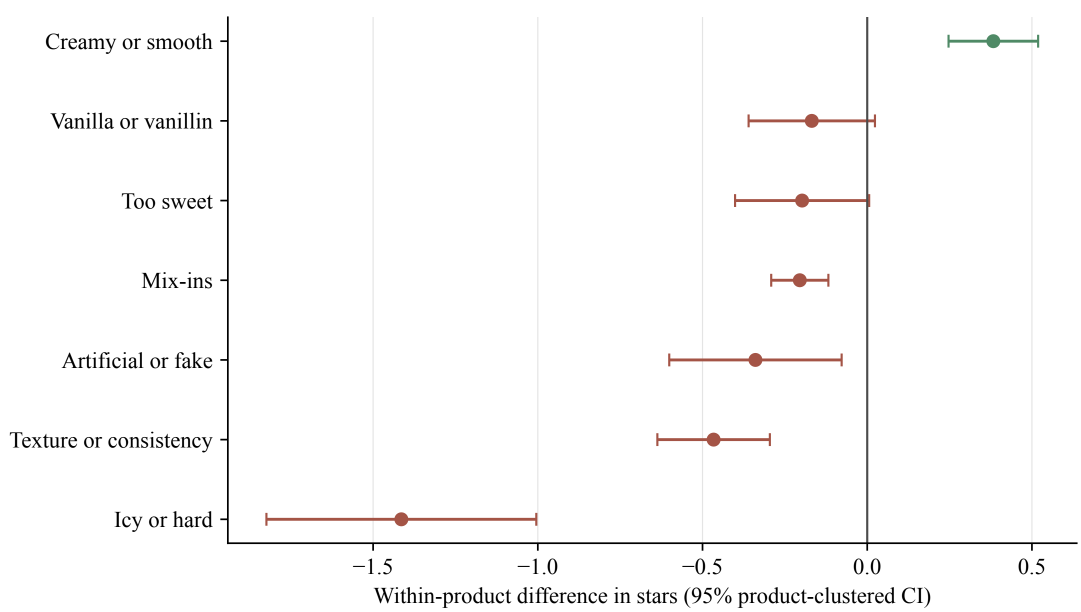

Figure 5: Within-product review-theme coefficients with 95% product-clustered confidence intervals. Color shows the direction of the estimate, not proof of cause.

4.6 Sensitivity checks

The four composition terms remain jointly nonsignificant with raw ratings ($p = 0.609$), EB prior strengths of 5 or 50 ($p = 0.528$ and $p = 0.411$), extra plain/reformulated indicators ($p = 0.738$), no brand indicators ($p = 0.818$), and only the 24 near-exact matches ($p = 0.635$). None of the versions I tested turns the main result into a stable curve.

Table 7: Sensitivity of the quadratic composition test

Specification	N	R^2	Adjusted R^2	Joint p
Primary EB, prior m=20	37	0.141	-0.066	0.438
Raw review mean	37	0.100	-0.118	0.609
EB, prior m=5	37	0.117	-0.097	0.528
EB, prior m=50	37	0.163	-0.039	0.411
Add plain/reformulated indicators	37	0.156	-0.125	0.738
No brand indicators	37	0.046	-0.073	0.818
Near-exact ingredient matches only	24	0.328	0.034	0.635

Notes: The joint p value tests fat, fat squared, sugars, and sugars squared. Unless noted, models include brand and the 37 accepted matches.

5 Discussion

5.1 The label is useful, but it is not a taste forecast

The main result is straightforward: across these four brands, fat and total sugars do not form a reliable rating curve. Ingredient-list length and the tested ingredient flags are also unstable, and composition models perform worse on held-out products than a mean-only guess. The null result is less dramatic than discovering the mathematically perfect scoop, but it is the honest result.

This does not mean fat and sugar are irrelevant. Controlled experiments ask what happens when researchers change chosen parts of a formula under similar conditions. This study asks whether labels explain ratings across companies that also change air, stabilizers, flavors, processing, storage, branding, and mix-ins. Guinard et al. [1] found a curved response under controlled formulations, while Rolon et al. [4] showed how a fat replacer can preserve acceptance. Both findings can be true at the same time as my weak cross-product coefficients.

The wide confidence intervals also matter. For sugar, the linear interval stretches from about -0.17 to $+0.72$ rating points per 5.25 g/100 g. The data do not prove that products are identical. The stronger conclusion is practical: the label variables do not predict new products well in this dataset.

5.2 Texture gives the clearest clue

Review-language differences are much bigger than the composition coefficients. Creamy and smooth words go with higher stars, while icy, hard, artificial, and general texture words go with lower stars. This lines up with sensory research showing that texture and flavor attributes help drive acceptance of commercial vanilla ice cream [3].

The word models still do not prove that one texture problem causes exactly a one-star change. A bad rating may make someone look for flaws, and a simple dictionary can misunderstand “not icy.” Still, the pattern suggests that future datasets should measure the physical spoon experience—hardness, ice-crystal size, melting, density, and serving temperature—instead of asking a nutrition label to do every job.

5.3 Why ingredient lists are weak numbers

An ingredient list tells shoppers what is present and gives a rough order by weight. It does not give exact amounts. “Natural flavors” can describe different flavor systems, and a tiny amount of stabilizer can interact with freezing and storage. Ingredient count also changes when a company expands subingredients or adds a complicated cookie swirl. This explains why the ingredient flags should not become buying rules.

Egg is a good example. It may represent a custard-style base, a richer brand family, or a different process. Its positive strict-sample estimate disappeared after correction and reversed in the broad sample. That is what confounding looks like in real data: the ingredient arrives with a whole crowd of other differences.

5.4 Limitations

This study has several limits:

1. *The effective sample is small.* There are 37 independent nutrition-matched products, not 5,793 independent recipes. Curves and brand controls use degrees of freedom quickly.

2. *Formulas come as bundles.* Nutrients, flavors, air, stabilizers, processing, storage, marketing, and price are chosen together. Brand controls cannot remove every missing product difference.

3. *The category is mixed.* Only 13 products are plain vanilla. The rest include gelato, modified products, other flavors, and mix-ins.

4. *Product matching is not perfect.* The review archive has no UPCs. The audit is conservative, and the near-exact-only analysis agrees, but some uncertainty remains.

5. *Time still matters.* The USDA file matches the archive endpoint, not every earlier review. Some recipes may have changed.

6. *Reviewers select themselves.* Manufacturer-site reviewers may not represent normal shoppers, and the star distribution is unusually positive [2].

7. *Keyword dictionaries are blunt tools.* They can miss sarcasm, negation, and context. The theme findings are descriptive.

8. *The sample covers four U.S. brands.* It cannot stand in for every store brand, small producer, country, or current formula.

5.5 A better next study

A stronger project would track exact UPCs across stores and weeks. It would save price, sales, package mass, package volume, promotions, stock status, and review dates. Price should be calculated per 100 grams and per 100 milliliters because ice cream contains air, so a pint and a pound are not interchangeable.

It would also measure overrun or density, ice-crystal size, melting rate, serving temperature, hardness, sweetness, and fat structure. Plain vanilla products should be separated from mix-in products before the model is fit. A blinded tasting panel could then test whether the online patterns match actual sensory ratings. Finally, the analysis should reserve whole brands or stores for external testing so a model has to prove itself somewhere new.

The best design would combine market data with an experiment. A controlled set of formulas could vary fat, sugar, air, and stabilizers over realistic ranges. The market data could then test whether those mechanisms survive the journey from the lab to the freezer aisle.

6 Conclusion

I connected 5,793 public ratings for 45 spoonable vanilla-labeled products to ingredient information and 37 matched USDA nutrition records. The curved fat-and-sugar model is not statistically reliable and predicts new products worse than a mean-only baseline. Ingredient count and common ingredient flags are also unstable after correction and broader checking.

The clearest pattern comes from what reviewers say. Creamy and smooth language goes with higher ratings; icy or hard texture, artificial taste, and general texture mentions go with lower ratings within the same product. Nutrition labels remain useful for nutrition. They just are not complete sensory models. The freezer, it turns out, keeps some of its secrets.

Data and Code Availability

The accompanying Overleaf bundle includes the complete LaTeX source, vector figures, generated tables, a 45-product analysis CSV, the nutrition-match audit, aggregate model results, and both Python scripts. It does not redistribute raw review prose. The original archives can be downloaded from Kaggle and USDA [5, 6]. Their recorded SHA-256 hashes are

- Kaggle archive:
  39745c0ef4932dd8d2fd3927661c06e14bae2216f779c860da0181a64f230304;

- USDA Branded Foods archive:
  ac9110e38532d8396b1ed893a270351bd2cdea69d62a66efa481e69e14ceb62c.

The Kaggle release is marked CC0, but this bundle still avoids copying individual review sentences. It distributes ratings, derived theme indicators, and product-level summaries instead.

References

[1] J.-X. Guinard, C. Zoumas-Morse, L. Mori, D. Panyam, and A. Kilara. Effect of Sugar and Fat on the Acceptability of Vanilla Ice Cream. *Journal of Dairy Science*, 79(11):1922–1927, 1996. doi: 10.3168/jds.S0022-0302(96)76561-X.

[2] Nan Hu, Paul A. Pavlou, and Jie Zhang. On Self-Selection Biases in Online Product Reviews. *MIS Quarterly*, 41(2):449–471, 2017. doi: 10.25300/MISQ/2017/41.2.06.

[3] Han Sup Kwak, Jean-François Meullenet, and Youngseung Lee. Sensory Profile, Consumer Acceptance and Driving Sensory Attributes for Commercial Vanilla Ice Creams Marketed in the United States. *International Journal of Dairy Technology*, 69(3):346–355, 2016. doi: 10.1111/1471-0307.12314.

[4] M. Laura Rolon, Alyssa J. Bakke, John N. Coupland, John E. Hayes, and Robert F. Roberts. Effect of Fat Content on the Physical Properties and Consumer Acceptability of Vanilla Ice Cream. *Journal of Dairy Science*, 100(7):5217–5227, 2017. doi: 10.3168/jds.2016-12379.

[5] Tyson. Ice Cream Dataset: Flavors and Reviews from Ben & Jerry's, Häagen-Dazs, and More. Kaggle, version 3, 2020. URL https://www.kaggle.com/datasets/tysonpo/ice-cream-dataset. CC0 1.0 Public Domain; updated October 4, 2020; accessed August 26, 2026.

[6] U.S. Department of Agriculture, Agricultural Research Service. FoodData Central: Branded Foods, October 2020 (CSV). FoodData Central data archive, 2020. URL https://fdc.nal.usda.gov/fdc-datasets/FoodData_Central_branded_food_csv_2020-10-30.zip. Release date October 30, 2020; accessed August 26, 2026.

[7] U.S. Food and Drug Administration. Ice Cream and Frozen Custard. 21 C.F.R. § 135.110, 2026. URL https://www.ecfr.gov/current/title-21/chapter-I/subchapter-B/part-135/subpart-B/section-135.110. Electronic Code of Federal Regulations; accessed August 26, 2026.

A Product Inventory

Table 8 lists every strict-sample product and the main variables used in the paper. A dash for fat or sugar marks one of the eight products left out of the nutrition regressions. Names are historical archive labels and some products may no longer be sold.

Table 8: Product inventory for the strict spoonable vanilla sample

Brand	Product	Reviews	EB rating	Fat	Sugars
Ben & Jerry's	Americone Dream®	370	4.71	15.22	24.64
Ben & Jerry's	Brownie Batter Core	108	3.61	13.38	24.65
Ben & Jerry's	Chocolate Chip Cookie Dough	978	4.55	14.49	23.91
Ben & Jerry's	Chubby Hubby®	118	4.28	18.92	23.65
Ben & Jerry's	Everything But The...®	72	3.89	17.69	25.17
Ben & Jerry's	Half Baked®	884	4.69	13.48	25.53
Ben & Jerry's	Ice Cream Sammie	31	4.68	16.67	23.53
Ben & Jerry's	Milk & Cookies	104	4.66	16.30	22.96
Ben & Jerry's	Red, White & Blueberry	39	3.70	12.69	18.66
Ben & Jerry's	Salted Caramel Almond	23	4.38	16.20	24.65
Ben & Jerry's	Vanilla	10	4.38	14.69	18.88
Ben & Jerry's	Vanilla Caramel Fudge	29	4.54	13.82	24.34
Ben & Jerry's	Wake & " No Bake " Cookie Dough Core	68	3.48	18.25	21.90
Breyers	CarbSmart™ Vanilla	157	3.99	7.69	5.13
Breyers	Cherry Vanilla	40	3.42	–	–
Breyers	Delights Vanilla Bean	167	4.18	–	–
Breyers	Extra Creamy Vanilla	59	4.39	5.56	20.99
Breyers	French Vanilla	101	4.49	11.49	20.69
Breyers	Homemade Vanilla	98	4.40	11.49	20.69
Breyers	Lactose Free Vanilla	60	3.99	5.17	21.84
Breyers	Natural Vanilla	467	4.12	10.23	21.59
Breyers	No Sugar Added Vanilla	18	4.30	5.41	6.76
Breyers	No Sugar Added Vanilla Chocolate Strawberry	34	4.29	5.41	5.41
Breyers	Non-Dairy Vanilla Peanut Butter	116	4.73	–	–
Breyers	Vanilla Caramel	37	4.20	–	–
Breyers	Vanilla Caramel Gelato Indulgences	85	4.43	7.62	23.81
Breyers	Vanilla Chocolate	95	4.50	10.34	21.84
Breyers	Vanilla Chocolate Strawberry	74	4.45	9.09	21.59
Breyers	Vanilla Fudge Twirl	22	4.08	–	–
Haagen-Dazs	Bourbon Vanilla Bean Truffle Ice Cream	89	4.06	15.15	25.76

Continued on next page

Table 8 (continued)

Brand	Product	Reviews	EB rating	Fat	Sugars
Haagen-Dazs	Cherry Vanilla Ice Cream	26	4.18	14.06	21.88
Haagen-Dazs	Coffee Vanilla Chocolate TRIO CRISPY LAYERS	48	4.32	20.33	21.95
Haagen-Dazs	Vanilla Bean Ice Cream	115	4.49	16.03	24.43
Haagen-Dazs	Vanilla Blackberry Chocolate TRIO CRISPY LAYERS	74	4.26	–	–
Haagen-Dazs	Vanilla Caramel White Chocolate TRIO CRISPY LAYERS	32	4.68	18.27	24.04
Haagen-Dazs	Vanilla Chocolate Chip Ice Cream	46	4.43	18.48	21.74
Haagen-Dazs	Vanilla Ice Cream	228	3.06	16.28	18.60
Haagen-Dazs	Vanilla Swiss Almond Ice Cream	129	4.33	–	–
Talenti	BLACK RASPBERRY VANILLA PARFAIT	65	4.41	10.91	23.64
Talenti	MADAGASCAN VANILLA BEAN GELATO	198	3.91	10.16	23.44
Talenti	ORGANIC OAK-AGED VANILLA GELATO	19	4.50	8.53	23.26
Talenti	PEANUT BUTTER VANILLA FUDGE	38	4.46	16.24	29.06
Talenti	VANILLA BLUEBERRY CRUMBLE GELATO	70	4.01	9.77	25.56
Talenti	VANILLA CARAMEL SWIRL GELATO	38	3.82	9.02	25.56
Talenti	VANILLA FUDGE COOKIE	114	4.29	–	–

B Matching and Feature Rules

B.1 How candidate matches were restricted

USDA candidates were limited to verified brand-owner names for the four companies and to U.S. records that could reasonably describe ice cream, gelato, frozen dairy dessert, ice milk, or a closely related item. Ingredient text was lowercased and stripped of punctuation and common function words. The matching features used both single words and adjacent word pairs. Product-name tokens left out brand names, numbers, package terms, and generic words such as *ice*, *cream*, *gelato*, and *dessert*.

The audit includes the best and second-best FoodData Central IDs, descriptions, ingredient similarity, name similarity, combined scores, threshold decisions, manual notes, and final inclusion. Complete ingredient strings remain available in the public source archives and are not repeated in the delivered audit.

B.2 Ingredient dictionaries

The ingredient indicators use these case-insensitive terms:

- *Gum/stabilizer:* guar gum, locust bean gum, carob bean gum, cellulose gum, xanthan gum, tara gum, or carrageenan.
- *Emulsifier:* mono- and diglycerides; monoglycerides; diglycerides; polysorbate; or lecithin.
- *Corn syrup:* the phrase *corn syrup*; high-fructose corn syrup is tracked separately.
- *Egg:* egg, eggs, egg yolk, or egg yolks.
- *Vanilla bean:* vanilla bean, vanilla beans, or vanilla seeds.
- *Vanilla extract:* the phrase *vanilla extract*.
- *Alternative sweetener:* allulose, erythritol, maltitol, monk fruit, sorbitol, stevia, or sucralose.

The ingredient parser tracks parentheses. A comma, semicolon, or period counts as a boundary only when it is outside parentheses. That is why I call the result a *top-level ingredient count*, not molecular complexity.

C Extra Notes on Robustness

The leave-one-brand-out test has only four folds, so it is a stress test rather than a complete market prediction score. Its quadratic RMSE of 1.840 comes from extrapolating outside the training brands' composition range. This is exactly the kind of warning cross-validation is supposed to catch.

EB shrinkage is modest for a normal-sized product. At the median of 72 reviews, the score gives the observed product ratings a weight of $72/(72 + 20) = 0.783$ and the platform distribution a weight of 0.217. At the maximum of 978 reviews, the platform weight is only 0.020. Changing this prior did not change the conclusion.

Accepted matches and uncertain matches have similar means, but eight excluded products are too few for a convincing equivalence test. The conservative approach trades some sample size for a lower chance of running a very precise regression on the wrong nutrition label.

D Reproducibility Manifest

The project is arranged for direct Overleaf compilation:

- main.tex is the paper, and references.bib contains the seven sources.
- figures/ contains five vector PDF figures.
- tables/ contains generated LaTeX tables and the complete product inventory.
- data/ contains derived product data, the match audit, coefficient results, cross-validation predictions, and a JSON result manifest.

- code/build_kaggle_usda_dataset.py constructs the matched data from the source archives.
- code/run_final_analysis.py fits the models and recreates all tables and figures.
- download_data.sh records source URLs, and requirements.txt records tested package versions.

The analysis is deterministic, so it does not need a random seed. Re-running it on byte-identical inputs gives the same numbers. PDF metadata and compression may still differ by compiler and date.

AI Assistance Disclosure

Artificial intelligence tools were used solely to improve clarity, organization, and grammatical precision of the written manuscript. All research design, mathematical modeling, data analysis, code implementation, experimental validation, and scientific conclusions were developed and executed independently by the author.