– Independent Work Report Fall 2021 –

Cannabis Analysis: Applying Statistical and Linguistic Methods to Understand Strain Names

Henrique Schechter Vera
Advisers: Professor Christiane D. Fellbaum [IW 02], Dr. Jérémie Lumbroso

January 10, 2022

Abstract

This paper details the extraction and analysis of cannabis strain data found in cannabis use and education website Leafly. The analysis applies regular statistical methods and methods drawn from the field of Natural Language Processing (NLP) to find patterns and relations between information such as strain names, ratings, effects, flavors, and genealogical origins. The analysis aims to carve a path to quantitatively investigate the factors involved in cannabis strain names and cannabis strain success, by organizing publicly available data and conducting an initial exploratory analysis through the lens of NLP.

1. Motivation and Goal

Cannabis strain names, simply put, are bizarre. Some, at least, have clear origins. "Sweet Dreams" is named after its purported relaxing effects, "Strawberry Cough" due to its strawberry flavor, "Lavender" because of its aroma, and "Jack Herer" in honor of the cannabis rights activism of its namesake. However, most are more complicated, and some are simply strange; some names of the most popular strains are “Martian Candy”, “Obama Kush”, “XJ-13", “Ewok”, “AK-47”, "Alaskan Thunder Fuck" and “Stardawg” [4]. In fact, some strain names are even counter-intuitive, connoting negative or otherwise unappealing characteristics. For example, some denote inedible or otherwise unpleasant substances, such as "Original Glue" and "Sour Diesel", and some even evoke danger, such as "White Widow".

As a result of their unusual and unexpected names, understanding how cannabis strains came to be named is intrinsically interesting. This process cannot be purely random, and so searching for some sense in the chaos of strain names may result in insights related to cannabis' chemical properties, human behavior, or perhaps even something completely unforeseeable.

More importantly, finding patterns in cannabis names, especially when relating these to strain popularity metrics, is bound to yield important insights for coming up with new strain names. This problem is a particularly interesting case within the field of product naming because the cannabis industry is actively transitioning from illicit to legal throughout the United States: "more than two-thirds of US states have legalized medical cannabis", 18 of those having also legalized cannabis for recreational use [9]. This unique in-between legal position implies strain naming practices are unique, and as growers, consumers, industry practices, and public perception all change, naming practices are likely changing, too.

The problem is also more relevant than ever, in part because of the upsurge in legality. Sales are growing: they hit $20 billion in 2020, were on track to overcome $26 billion in 2021, and are projected to leap to $45.9 billion in 2025 [9]. Jobs are increasing: there were an estimated 321,000 full-time jobs in the cannabis industry in 2020, up from 234,700 just a year before [9]. Research into cannabis strain names, then, is also more significant than ever.

2. Background and Related Work

One significant reason for which this project is interesting is that little to no work has been done on the area of cannabis strain naming, despite the study of cannabis overall being being a well-established field. Searches on the Association for Computational Linguistics' anthology for "cannabis strain names", "cannabis strain", and "strain names" yield some cannabis-related papers and projects, but no relevant results. Similar searches on JSTOR yield the same results.

One relevant paper published in the journal *Cannabis and Cannabinoid Research* attempted to "determine the actual levels of chemical diversity represented in 2662 samples of Cannabis flower collected between January 2016 and June of 2017 in Nevada" [6]. Researchers measured chemical

profile data and used it to cluster cannabis samples with different breeder-reported names using principal component analysis (PCA). It turns out thousands of samples drawn from 396 differently named strains cluster best into only 2-3 well-defined groups, depending on the exact definition of chemical profile. See in Figure 1, for example, how only two clusters are optimal when using both terpenoids and cannabinoids in defining the chemical profile (results were very similar when using each type of chemical individually). Thus, the paper concluded that cannabis strain names overestimate the chemical and genetic diversity of strains "and do not inform patients regarding chemical properties". Chemical makeup being unable to explain differences in strains' names begs the question of what could. Perhaps use effects such as "euphoric" or "giggly", which occur due to only slight chemical differences? Or perhaps only perceived differences in strains, communicated implicitly by connecting strains to well-established parent strains? These results have even caused people to call for either scratching cannabis strain names altogether or enforcing regulations with regards to names, further motivating this project [5].

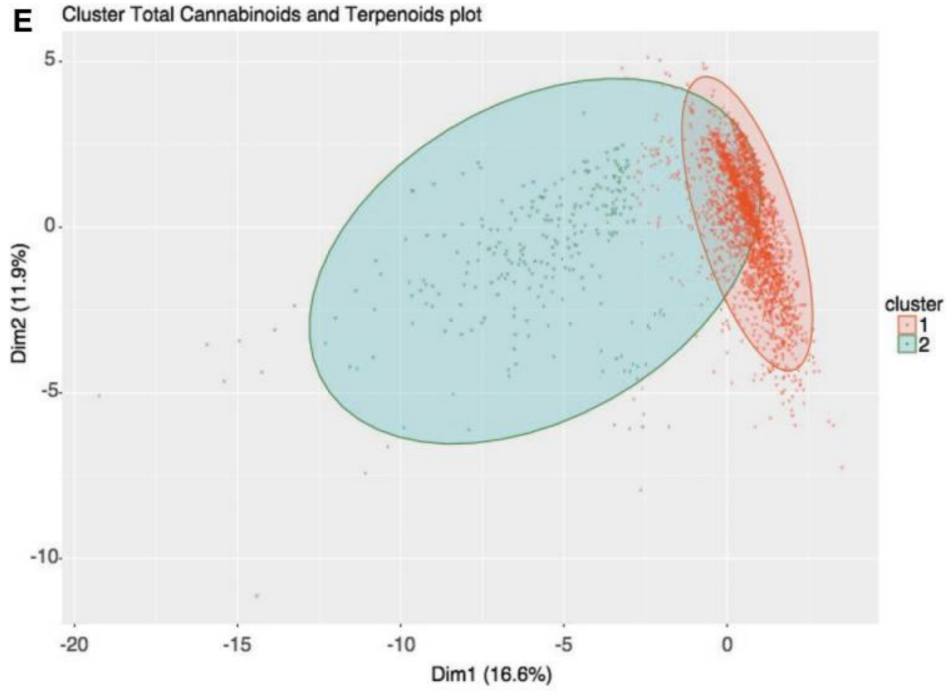

Figure 1: Optimal clustering using cannabinoid and terpenoid data.

Another relevant paper published in the journal *JAMA Psychiatry* seeked to compare cannabis

flowers with differing tetrahydrocannabinol (THC) concentrations using biobehavioral research methods [1]. Researchers found that THC concentration of cannabis products were unrelated to user intoxication. See in Figure 2 how the dotted lines, representing the cannabis products with lowest THC concentrations, actually yielded the highest intoxication levels. This lies in stark contrast to public perception: "when it comes to moving product on the legal recreational [cannabis] market, only two numbers matter: the list price, and the THC content" [7]. If differences in consumption effects also cannot be explained by strains' chemical composition, then could it be that these purported effects–and perhaps even flavors and alleged health benefits–are to some extent placebo and thus mentally induced by strains' names?

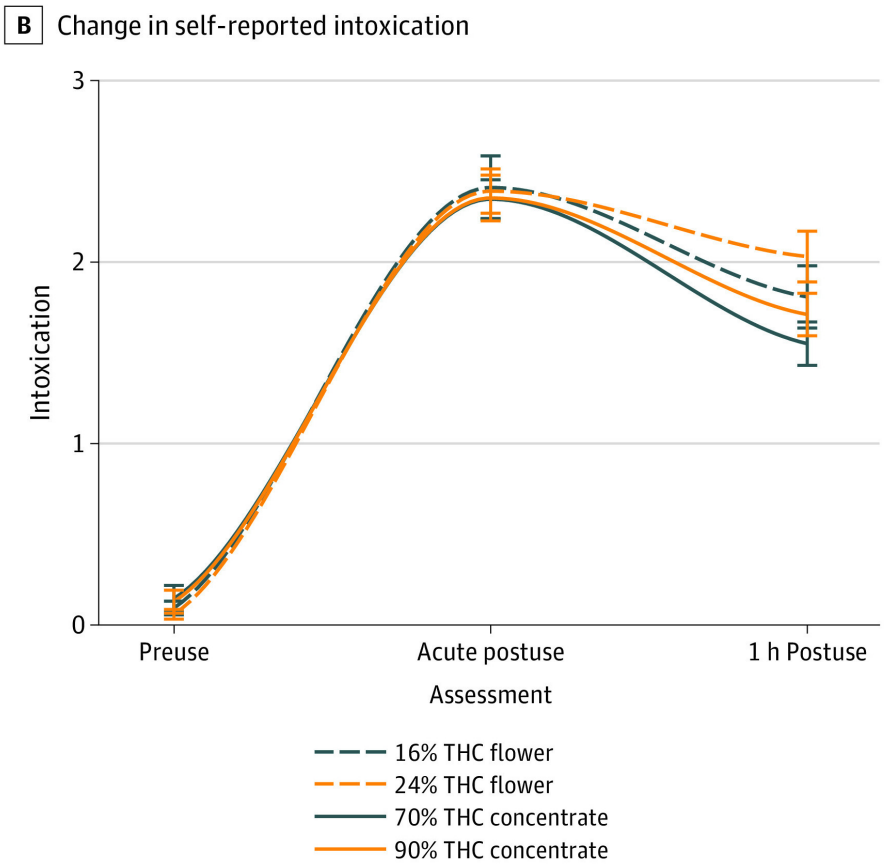

Figure 2: Change in self-reported intoxication between products with different THC concentrations.

Both of these results yield interesting motivating questions. But more importantly, they inspire and justify the approach we take in this project, described next.

3. Approach

Despite both strain names and use effects apparently are not attributable to chemical properties, users continue to differentiate between strains and report wildly different consumption experiences. We need a new approach to understanding (differences in) strain properties like consumption effects and names. The lack of chemical explainability suggests this might be rather psychological and due to perceived differences in how strains are portrayed or marketed. As such, we choose to adopt a linguistic lens with which to tackle the questions.

Moreover, thanks to the ongoing legalization of cannabis use, cannabis suppliers and websites can grow rather than stay small and hidden. In combination with the digital revolution, the result is that there is more and more data about cannabis strains available online. This allows for computational analysis, which in combination with our linguistic lens naturally invites the use of techniques drawn from Natural Language Processing (NLP), by definition an umbrella term for computational processing of natural human language, such as sentiment analysis and semantic embeddings.

However, current publicly available datasets are lacking mostly in that they either (1) focus not on attributes of cannabis strains themselves but rather on legal and business attributes of the cannabis industry at large, or (2) are not large enough, both in amounts of fields and in amounts of strains. As such, we chose to create our own dataset using data found in Leafly. We chose Leafly for its size of more than 125 million yearly visitors, 1.3 million product reviews, and 5,000 strains [3]. Compare our resulting dataset to the best I could find online: 5774 strains, instead of approximately 800, and 71 total fields, instead of approximately a half-dozen [2].

As a result, our approach is novel in three ways:

1. We investigate questions and hypotheses that have been left unexplored altogether; this field of research is new.
2. We apply natural language processing (NLP) and computational linguistics (CL) techniques for analysis.

3. We use our own dataset, which contains the newest, largest, and most comprehensive strain data available.

4. Implementation

4.1. Data Scraping

Feel free to access our complete datasets right here.

4.1.1. Strain attributes

Scraping data from Leafly used to be easier in the past. Each of the thousands of cannabis strains in the database used to simply have been assigned increasing numbers, starting from 1, that identified the strain. Each strain then had a corresponding page which one could reach by adding that number to the end of a fixed URL address (page address). For example, "leafly.com/strains/1" corresponded to the first strain in their database, "leafly.com/strains/1" to the second, and so on. Recently, however, the extensions to be added to the URL address "leafly.com/strains/" (which take the name of "slugs") were recently changed into more meaningful textual representations of the strains, *almost* always matching strains' names. For example, the strain "Sorcerer's Apprentice" can be found at "leafly.com/strains/sorcerers-apprentice". To access each strain's page, then, we first began by going through Leafly's list of strains, accessible at URLs "leafly.com/strains?page=1" all the way to "leafly.com/strains?page=182", and storing these slugs from the hyperlinks that led to each strain.

After this, we would be able to extract the data from each individual strain's page. A lot of time was spent figuring this out, since much of the relevant information was displayed in bars (strain potency), graphs (children and parent strains), and other complicated web page elements. However, careful examination of the source code of the pages that list all strains revealed that *all* remotely relevant information about the strains that Leafly collects, only some of which is actually visible in strains' individual pages, are actually all conveniently available in that source code in JSON format, despite not being utilized in the pages listing all strains. As a result, our scraping must only iterate through 183 pages of strain lists, each listing at most 30 strains, instead of nearly 6,000 individual

pages. Because querying a page and extracting its source code is the most time-consuming part of the scraping itself, our scraping turned out to be extremely efficient.

Now, we discuss more specific implementation details. We performed our scraping in Google Research's Colab platform, so as to (1) be able to more easily store things in the cloud through Google Drive, and (2) because it is built for Python, which contains many relevant libraries for things such as scraping and NLP. Ultimately, then, the way we scrape for information about strains is, for each of the 183 pages listing strains:

1. We query the source code using the 'requests' library.
2. We extract the information about the strains listed in the given page (found easily because it is consistently found in the same location in the source code).
3. We use the 'json' library to convert the new data into a dictionary.
4. We extract our ongoing data thus far from a file stored in our Google Drive using the 'shelve' library.
5. We add our new data to the ongoing data as a new entry in a list.
6. We update the Google Drive file with the newly updated total data.

Note that constantly updating our database file using 'shelve' ensures that our progress is saved in case the process is cut out due to network or software errors, or in case we cannot afford to complete the entire scraping process in one go (it took a very long time).

In the end, we are left with a list of 5774 dictionary objects, one for each strain, and each with 71 different fields. See an example in Figure 3. For convenience, I also selected and list below the set of attributes I found to be most potentially useful or relevant:

- Aliases
- Average rating (and rating count)
- Awards
- Cannabinoid percentages (CBC, CBD, CBG, THC, THCV)
- Category or phenotype (sativa, indica, hybrid, edible)
- Chemotype (THC or CBD dominant, or balanced)

- Energizing score and highness/THC score
- Terpene (aromatic chemicals) concentrations
- Children strains, parent strains
- Scores for each:
  - Condition (e.g. anxiety)
  - Effect (e.g. relaxed)
  - Negative effect (e.g. sore throat)
  - Flavor (e.g. earthy)
  - Symptom (e.g. lack of appetite)
- Growth information (difficulty, yield, height, flowering days, etc.)
- Similar strains
- Total followers
- Written description

```
{
    'slug':                'jet-fuel',
    'id':                  118799,
    'aka':                 'Jet Fuel OG, G6, Jet Fuel G6,
                           Jet Fuel Kush, G6 Kush',
    'articleTotalCount':   5,
    'articlesAvailable':   True,
    'averageRating':       4.523809523809524,
    'award':               {'blurb': None,
                           'imageUrl': None},
                  […]
    'trending':            False,
    'videoUrl':            None
}
```

Figure 3: Representation of a strain object, with truncated information, for strain "Jet Fuel".

4.1.2. Strain reviews

The former approach was sufficient to scrape data on almost all strain attributes, except one: reviews. These are particularly tricky because (1) they are not found in raw page source content, (2) they are found separately altogether, in strains' review pages, and (3) there are a lot of them, and they take up a lot of memory.

We solve problem (1) first. For a given strain's slug *slug* and the number *pagenum* of one of its review pages, we use the page address "https://www.leafly.com/strains/*slug*/reviews?page=*pagenum*". Since this page's source content alone is insufficient, we use the 'HTMLSession' module from the 'requests_html' library to render the page's HTML with its JavaScript. Then, we can use the 'BeautifulSoup' library to find the review elements.

One issue with this process is that it turned out to be unreliable, sometimes yielding no reviews despite a page being populated with reviews. Thus, we attempt this process repeatedly until we get a non-zero amount of reviews. However, now we must know whether we expect a page to have reviews or not. Thus, we must figure out how many pages of reviews each strain has. We do this by visiting the first page of reviews for each strain and following the same process as above, except instead of searching for the review elements using 'BeautifulSoup', we search for the 'max page number' element. We stored these page totals in a small database file, again using the 'shelve' library.

The final issue was space. The HTMLSession was incompatible with Colab, which is constantly running on an event loop, and so scraping the reviews was done locally on the PyCharm IDE (integrated development environment). However, we needed cloud storage more than ever, given Leafly reports having 1.3 million user reviews. The solution was to use the GoogleDrive (in conjunction with the required GoogleAuth module) from the 'PyDrive' library. After some credential setup through Google Cloud Platform, this allows us to directly upload to our Google Drive.

Thus, ultimately our process for scraping reviews was very similar to that of scraping other strain attributes:

1. Access our lists of page totals and URL slugs for all strains.
2. For each strain, iterate over its page numbers. For each:
    (a) Scrape reviews in the corresponding page, as described above.
    (b) Retrieve our data thus far, stored in the form of a dictionary where each strain slug is a key, using the 'shelve' library.
    (c) Add to list of reviews in the dictionary entry corresponding to the strain.

(d) Update our stored ongoing data using ‘shelve’.

One important detail is that ‘shelve’ files are themselves kept as dictionaries, storing a dictionary within one, which itself has values that are lists, entails a data structure that is too recursive by default: the program will yield an error. Thus, it is important to set the recursion limit to a higher value using the ‘sys’ library.

Note that since each page of reviews holds at most 8 of a stain’s total reviews, there are a lot of pages to scrape; moreover, each page must be rendered by JavaScript before being scraped; moreover, the data must be sent through a server to be stored in the cloud. As a result, this process is very lengthy. As a result, before the review dataset could be completed, Leafly was restructured. Leafly has begun to require Captcha verification to use, and, as of the final version of this paper, continues to have errors in displaying data (e.g. no strains at all can be found in the list of all strains).

4.1.3. Categories

As part of a subsequently explained test (section **5.3**), we created two relevant lists of categories that may also prove useful for future analysis. The first is simply a list of names of fruits. The second is a list of names of colors, grouped into thirteen main "shades". For example, "cardinal", "salmon", and "maroon" are all under "red".

Here, we also used the ‘requests’ library to get the page’s source content, the pages this time being from Wikipedia pages [8] [10]. We then use the ‘BeautifulSoup’ library to parse the HTML and find the elements corresponding to fruits and colors, respectively. We then store our results also using the ‘shelve’ library; the fruits are stored as a list, and the colors as a dictionary with keys being the names of the main shades (blue, red, green, etc.) and the values being a list of colors that fall under that broad category. We also preprocess color names by substituting parenthesized expressions with the ‘re’ regular expression library (and of course, as always with textual data, trimming trailing and beginning whitespace and converting to lower case).

4.2. General Testing Methodology

There are two critical aspects of our hypothesis testing methodology to discuss: the statistical methods used in actually testing our hypotheses, and the strain attribute we used as the popularity metric (since strain popularity was relevant to many hypotheses).

First, our statistical methods. We began by running various regression analyses between strain attributes that were either already quantitative or that we computed or quantified ourselves. This seems to be the most intuitive approach to take when trying to find relations between data, as regressions by definition attempt to find relationships between variables. However, we soon found these tests to yield extremely statistically insignificant results (*very* low p-values). This was in spite of the large amount of tests we conducted–in fact, we conducted so many tests that one concern going in was that we would get false positive results out of mere chance. We conjectured that this was due to (1) correlations already expected to be weak because many external forces (that are unaccounted for here) come into play with strain naming and strain popularity, and (2) we had a large sample size of 5774 strains, and each strain introduced variance, such that their combined variance was too large to yield statistically significant results with regular regression analysis. Note that we performed linear regression analyses specifically, all using the 'LinearRegression' module of the 'sklearn' library (which required us to represent our data using 'numpy' arrays), because it would be unreasonable to expect complex non-linear relationships between the variables we were testing, though it might have been worthwhile to explore other regression tests had we not found this issue with regression tests as a whole.

As a result, we pivoted our approach for future questions. To try to account for the variance that thousands of strains introduced, we instead separated strains into two groups based on one attribute (the independent variable), and ran t-tests for differences in the group means of another attribute (the dependent variable). The decrease in variance is two-fold. The binary categorization accounts for variance in the independent variable, as small fluctuations in the independent variable would not be large enough to change which of the two groups a strain is put in. More importantly, taking the mean of the groups accounts for variance in the dependent variable, as each group would have

more than enough strains to have a representative sample. This change of direction also meant reformulating our questions accordingly, to accommodate this type of hypothesis test. All T-tests were computed using the ‘ttest_ind’ function from the ‘scipy’ library’s ‘stats’ module, which offers both one-sided and two-sided T-tests.

The second important aspect to discuss is the heuristic we use for strain popularity. First, we simply used average strain ratings. However, this is limited in a few ways. First of all, many strains do not have many reviews, implying high variance in their average rating. Second, strain ratings are not very evenly distributed between 1 and 5, tending to be rather high. This could be because strain consumption tends to generally be enjoyable, or because those leaving reviews tend to self-selectively be happy consumers. We dealt with this by defining "popular" strains to be those rated above 4.7, as this is almost exactly the 75th percentile of average ratings, but it might be worthwhile exploring normalization methods instead, especially for tests where popularity was the dependent variable.

We then ran each test or analysis again using amount of ratings as the popularity metric instead. This measure is inferior to the former, of course, because it encompasses strain quality only implicitly, since better strains are more likely to be sold frequently and thus receive many reviews. However, it is superior in a few ways. First, it is not subject to variance as a result of small sample sizes. Strains with few reviews are out there competing in the market and simply are not as popular. Second, it is less subjective. Strain ratings are human attempts at quantifying their experience. Market success and sales volumes, which is reflected in review amounts, are less qualitative. However, this measure also shares the problem that average rating had of being a skewed distribution (though positively skewed, rather than negatively skewed). As such, we also defined "popular" when using amount of ratings as above the 75th percentile, which is those with 29 reviews or more.

We note that it might be interesting combining the two measures for a more comprehensive popularity heuristic, perhaps also scaling the weight of ratings by the amount of reviews. This is especially true since the two measures of popularity we used often yielded highly contradictory

results in the same tests, and so are both evidently limited as individual indicators of popularity.

5. Analyses

Within each motivating question's subsection, we discuss both implementation and results.

5.1. What strain attributes might help make them popular?

Here, our goal was to find possible predictors of strain popularity, so that we could determine whether chemical composition, consumption experiences, or marketing (namely strain names) influence consumers most. The intent was to see whether names really were worthwhile to look at if creating popular strains is one's goal. We fit linear regression models with strain popularity as the dependent variable. The independent variables we tested were:

1. THC concentration.
2. User-reported effect scores. Effects here refers to feelings and experience upon consuming the strain, specifically 'Relaxed, 'Sleepy', 'Creative', 'Talkative', 'Euphoric', 'Energetic', 'Hungry', 'Giggly', 'Tingly', 'Happy', 'Focused', 'Aroused', and 'Uplifted'.
3. User-reported flavor scores. Flavors here refer to 47 different tastes and aromas like 'pungent', 'nutty', 'earthy', and 'grape'.
4. Sentiment values of strain name. These were computed using the 'SentimentIntensityAnalyzer' module of the 'vaderSentiment' (or 'VADER') library. VADER's sentiment analyzer was chosen intentionally. First, it is lexicon-based (i.e. considers each word individually). This is appropriate here because strain names are often groups of unrelated words (often even comically so) put together, so attempting to consider the names as existing phrases or in context of each other would likely produce less accurate scores. Second, it allows us to compute "compound" polarity scores, which combine scores for both positive and negative sentiment, boiling sentiment down to a single value and thus making analyses easier to carry out.

Results: We found only weak or insignificant relationships between the variables. R^2 values did not exceed a magnitude of 0.01. We fail to reject the null hypotheses that there are no significant

relationships between strain popularity and these strain attributes. See full results in the Appendix, in Table 4.

5.2. To what extent are strain names attributable to strain consumption effects (or vice versa)?

Sometimes, the influence of a strain's effects on its name are clear, as is the case with the "Laughing Buddha" strain having 'giggly' as its top user-reported effect, or example. However, most strains' names are *not* direct references to their most commonly reported effects. Still, it's possible that strains' names tend to generally reflect the sentiment of a strain's top reported effects. This is what we test here. Specifically, we compute compound polarity scores (as justified and explained in **4.2.2** above) for each strain's names and most commonly reported effect, and fit linear regression models with sentiment of strain names as the dependent variable and sentiment of strain effects as the independent variable (though a causal effect is possible the other way, too, since it's possible that users experience effects in part as placebos due to strain names).

Results: Our model yielded a coefficient of 0.025630 and an intercept of -0.00061733 with an R^2 value of 0.0023264. This coefficient of determination is extremely low, and so we fail to reject the null hypothesis that strain names are not attributable to strain consumption effects.

5.3. What strain consumption effects are the most conducive to popularity?

Many strain names signal the strain's consumption effects, as with "Laughing Buddha" inducing giggliness and "Green Crack" purportedly leaving users energetic. Thus, it would be useful to know which of these effects sell best, so producers can potentially choose names indicative of top-performing effects (with respect to strain popularity).

For a given effect E_i, we separate strains whose top reported effect is E_i from all other strains. We then conduct a two-sided T-test for differences in means of the popularity metric. We also compute the differences in mean values of the popularity metrics themselves, to see what the actual change in popularity is.

Results: Using average ratings, many effects yielded statistically significant results at an $\alpha = 0.05$

level: ‘sleepy’, ‘tingly’, ‘focused’, and ‘aroused’. However, the differences in means are minimal, with the largest being with the ‘aroused’ group, whose strains averaged a rating nearly 0.1 larger than that of other strains (on the 5-point scale).

Using review counts, many effects also yielded statistically significant results at an $\alpha = 0.05$ level: ‘sleepy’, ‘energetic’, ‘tingly’, ‘focused’, and ‘aroused’. Here, however, differences in means were actually very large, all in the magnitude of hundreds. Notably, while ‘sleepy’ and ‘energetic’ effects seem to positively contribute to review counts, strains whose top effects were ‘tingly’, ‘focused’, or ‘aroused’ actually had significantly *fewer* reviews than other strains. Intuitively, these results seem to make sense, as the former two effects are broad and thus more likely to be popularly appealing, while the latter three are specific and niche such that they are less likely to be.

As mentioned above, we *do* find that certain effects significantly contribute to strain popularity, since for each of these effects we reject the null hypothesis that strains that most commonly yield the given effect on average have the same popularity scores as other strains. See complete results in Table 1.

Effect	Difference in means	T-statistic	p-value
Relaxed	0.038728	1.7212	0.085331
	-8.0992	-0.24613	0.8056
Sleepy	-0.055583	-3.1411	0.0017009
	95.505	3.6951	0.0002241
Creative	-0.018064	-0.62371	0.53287
	31.182	0.73662	0.46142
Talkative	-0.0036803	-0.13688	0.89114
	-65.12	-1.6578	0.097466
Euphoric	-0.025055	-0.85154	0.39455
	8.4848	0.19726	0.84364
Energetic	-0.065308	-3.0728	0.0021418
	218.67	7.0909	1.6879e-12
Hungry	-0.060787	-2.3273	0.02002
	8.3714	0.21906	0.82662
Giggly	0.036334	1.4206	0.15554
	-60.715	-1.6243	0.10443
Tingly	0.045713	1.8085	0.070635
	-113.09	-3.0644	0.002202
Happy	0.060494	2.0305	0.042407
	-65.816	-1.5108	0.13095
Focused	-0.00090716	-0.035037	0.97205
	-93.862	-2.483	0.013087
Aroused	0.099128	3.754	0.00017764
	-116.88	-3.0255	0.002505
Uplifted	0.044411	1.5483	0.12168
	-73.396	-1.7508	0.080094

Table 1: Differences of means of popularity between strains with given top effect and the rest of strains (difference = mean with effect - mean of rest). Popularity is defined as average rating in white rows, amount of reviews in gray rows.

5.4. Do name patterns with categories influence strain popularity?

One pattern observed in strain names is that these will often contain words that clearly belong to one of many categories. Most commonly, many strains contain a color in their name, and many contain a fruit. Sometimes, multiple strains exist whose names are identical except for a change in the category word. For example, there is both a strain called "Blueberry Dream" as well as one called "Strawberry Dream", and there is both a strain called "Purple Kush" as well as one called "Green Kush". Sometimes the names are the same with the *type* of category having changed. For

example, besides the ‘Kush’ strains with colors, there are also ‘Kush’ strains with fruits, such as "Banana Kush" and "Orange Kush".

Here, we try to figure out which *type* of category is mot conducive to popularity (between fruits and colors) by running a two-sided difference of means T-test between mean popularity scores of strains with fruits in their name and strains without fruits in their name, and then doing the same with colors. We also test whether having category words at all is conducive to popularity, by running a two-sided difference of means T-test between mean popularity scores of strains with either a fruit or a color (or both) in their name and strains with neither in their name.

We manually created lists of categories for these two category types because these appear to be the most popular category types within strain names. Note that a less accurate but faster approach would entail using named-entity recognition (NER) algorithms to label words with categories and thus allow similar analyses to be performed for many other categories, such as animals.

We note that it would be interesting to find which specific categories are the most conducive to popularity, within each type, by running the same difference of means tests but between strains with those specific categories in their name and those without them. However, at least with our data, these results would not be very meaningful: there is an insignificantly small amount of strains, for example, which utilize the word "orange" specifically.

Results: Most results turned out to be statistically insignificant. The test that separates strains with either a fruit or a color in their name yielded very similar results to the test that only separated strains with fruits in their names (both significant), while the test that only separated strains with colors in their names had very different and very insignificant results. This makes it seem very likely that it was the fruit names that caused the statistical significance in the "fruit or color" test. Thus, we reject the null hypothesis that strains with no category words in their name have the same mean popularity as those with fruit or color category words, and we also reject the null hypothesis that strains with no category words in their name have the same mean popularity as those with fruit category words, both when using average rating as the popularity metric. However, counter-intuitively, we found that in both of these cases, the group with the higher mean average

rating was actually that without any category words in their name.

See full results in Table 2.

Category type	Means (contains word)	Means (no word)	Diff. of means	T-statistic	p-value
Fruit or color	3.4951	3.6421	-0.14696	-2.3197	0.020391
	66.172	67.652	-1.4806	-0.11861	0.90559
Fruit only	3.5052	3.6386	-0.13341	-2.0606	0.039391
	68.664	67.161	1.5023	0.11777	0.90625
Color only	3.4405	3.6193	-0.17877	-0.80121	0.42305
	47.227	67.636	-20.409	-0.46442	0.64237

Table 2: Differences of means of popularity between strains that contain a category word and those that do not, for different sets of category words (fruits, colors, and both). Note that difference = mean containing word - mean with no word. Popularity is defined as average rating in white rows, amount of reviews in gray rows.

5.5. Do names of popular strains share common characteristics?

Here, we are trying to see if there are patterns to popular strains' names. We investigate linguistic (lexical) traits; specifically, we examine part-of-speech distributions and sentiment values of strain names.

We separate all strains into two groups: popular (avg. rating > 4.7 or review count > 28) and unpopular. We then carry out two-sided difference of means T-tests on (a) names' compound polarity (sentiment) scores, on (b) proportion of adjectives in strains' names, and (c) proportion of verbs in strains' names. We choose adjectives and verbs specifically because they are the only parts of speech that occur with any meaningful amount of frequency in strain names (after nouns, of course).

We used the 'nltk' (Natural Language Toolkit) library here: the 'punkt' module for tokenizing and the 'averaged_perceptron_tagger' module for part-of-speech tagging.

Results: The tests yielded significant results for proportions of verbs and proportions of adjectives, but not for sentiment value. Tests with proportions of verbs had the smallest p-values and yielded significant results on both popularity metrics.

However, these results might not be too promising: for both proportions of adjectives and

proportions of verbs, the differences of means changed from positive to negative between popularity metrics, which *seems* to indicate our results are actually contradicting themselves. It may be the case that proportions of adjectives and verbs in strain names are already so small that even slight variance in their values is sufficient to create false positives in these tests.

See full results in Table 3.

Difference in means	Popular mean	Unpopular mean	Diff. of means	T-statistic	p-value
Sentiment value	0.0065748	0.0065564	1.8346e-05	0.0036272	0.99711
	0.011889	0.0047755	0.0071133	1.4493	0.1473
Proportion of verbs	0.017069	0.01113	0.0059396	2.3569	0.018462
	0.0058247	0.014731	-0.008906	-3.6437	0.00027114
Proportion of adjectives	0.039527	0.049565	-0.010038	-2.0218	0.043241
	0.050173	0.046279	0.0038938	0.80784	0.41922

Table 3: Differences of means of lexical properties of strain names between popular strains and other strains. Note that difference = popular mean - unpopular mean. Popularity is defined as average rating in white rows, amount of reviews in gray rows.

5.6. Does the biological genealogy of strains account for their popularity?

Most strains are results of breeding existing strains. Here, we seek to determine whether strains which have at least one popular parent score higher in popularity themselves (than those that do not have any popular parents). Again, we separate the strains into those with at least one popular parent (avg. rating > 4.7 or review count > 28) and those without any, and run one-sided difference of means T-tests on popularity values (the alternative hypothesis being that the group with children of popular parents will have a *greater* mean).

Results: Both popularity metrics yield very large differences in means with very high statistical significance.

Children with *at least one* parent with a rating above 4.7 had an average rating of 4.15, while children with *no* parents with a rating above 4.7 had an average rating of 3.60. The T-test for the difference between these means yielded a T-statistic of 3.8186 and a respective p-value of 0.00013559.

Children with *at least one* parent with a review count above 28 had an average review count of

16.489, while children with *no* parents with a review count above 28 had an average review count of 124.26. The T-test for the difference between these means yielded a T-statistic of 11.637 and a respective p-value of 5.9340e-31.

It might also be productive to investigate if and how this effect dwindles as one goes down the descendant line. For example, might this effect of increased popularity wane in strength as a linear function of how many ancestors away a strain is from a popular strain?

5.7. Does 'signaling' parent's name increase popularity retention?

We found that children of popular strains tend to be more popular themselves. Could it be the case that this effect is caused (or at least amplified) by signals in the children strains' names that suggest or communicate which strains they are descended from?

To answer this, we test to see whether strains with these signals tend to have popularity scores closer to those of the parents whose names they are signaling. We consider a signal to be a word that is present in a strain's parents' names. Thus, the question we are really asking is whether strains which share a word with one or both of their parents (e.g. "Bubba Kush", children of "OG Kush" and "The Bubba") have more similar ratings to those parents?

We do not neglect stopwords (commonly used words that are typically ignored during NLP tasks), as even words like "The" in "The Bubba" are significant when strain names are composed of such few words and tend not to contain stopwords. We first separate strains into those that share at least one word with either of their parents (this group definition is lenient because it is relatively infrequent), and those that do not share any. We then compute the average absolute difference in popularity scores between the child strain and its parent strains. We define that as its "distance" in popularity from its parents, and compute a one-sided difference of means T-test of popularity scores (the alternative hypothesis being that the group with children with signals will have a *smaller* mean).

Results: Using average rating, strains with at least one signal did indeed have a smaller "distance"

in popularity from their parents than those that had none, but only slightly so: the former group had an average distance of 0.48521 and the latter 0.52773. This result was very statistically insignificant, though, with the T-test yielding a T-statistic of -1.1416 and a corresponding p-value of 0.12687.

Using review counts, strains with at least one signal had a mean "distance" from their parents of 1854.1 reviews, while those with no signals had a mean "distance" from their parents of 1492.3. This result is even more insignificant: the T-test yielded a T-statistic of 4.5205 and a corresponding p-value of 0.99999.

Interestingly, using review counts, the effect is actually the opposite of that which we expected. Indeed, re-running the T-test using the opposite alternative hypothesis yields an extremely significant p-value of 3.2119e-06. Perhaps users are less inclined to write reviews for strains that they perceive as similar to strains they have already written reviews for (the popular parents). This result seems to contradict the intuitive explanation for our prior finding that children of popular parents tend to be popular themselves, which is that the children are popular by virtue of being recognizably related similar to their parents. Perhaps it is the chemical properties of the popular parents that make the kids popular, too, but children must not seem too obviously similar to their parents (e.g. by having signals in their names) so as to appear to be a unique strain and not a rip-off. This would also help explain the unexpected outcome in **Section 5.4** that strains with names that did *not* have category words tended to have higher average ratings.

There are two other related tests we did not carry out that might yield interesting results. First, it would be interesting to carry out this same test using not just exact word matches but also strength of semantic similarity to parents (using word embeddings). Second, it would be worthwhile to test, for strains that have a child *with* a shared word as well as a child *without one*, whether the former children have popularity scores more similar to those of the parent strain than the latter. In other words, this would test if children strains with "signals" in their names have popularity scores closer to their parents' than their "sibling" strains that do not have signals in their names.

6. Summary

6.1. Conclusion

We accomplished what we set out to do with this project. In seeking to investigate questions and hypotheses that have been thus far unexplored, we carried out several analyses and ran countless tests: this was a great initial exploratory analysis of cannabis strain names through a computational and linguistic lens, using techniques such as sentiment analysis and part-of-speech tagging as well as a new and more comprehensive dataset.

We had several statistically insignificant results, many unexpected, but here we summarize the statistically significant results we acquired. We found evidence supporting the ideas that:

1. Certain effects ('sleepy', 'energetic', 'tingly', 'focused', and 'aroused') significantly contribute to strain popularity. Contributions to review count were much larger than to average rating, but were often *negative* contributions (decreasing popularity).
2. Strains with names that contain fruits (or perhaps categories at large) tend to be *less* popular than those that do not.
3. Strain names with verbs tend to have higher average ratings but lower review counts. Strain names with adjectives tend to have lower average ratings.
4. Strains with at least one parent that is popular (75th percentile or higher in popularity) tend to have much higher average ratings and review counts (on average, 0.55 points higher in rating with 107.771 more reviews).
5. Opposite to what was expected, strains that signal their popular parents in their names tend to differ in popularity from their parents more than those that do not (using review counts).

The results may influence how cannabis sellers breed, market, and name new strains.

The first three reveal three ways in which they may do so such that they exploit tendencies for popularity. For example, they may breed strains to yield specific effects that tend to be related to popular strains, advertise strains to be particularly potent at evoking these effects (perhaps indirectly

through names reminiscent of effects, like "Northern Lights" being relaxing), or avoid fruits and adjectives when naming strains.

The latter two reveal potential for exploitation. Evidently, strains somehow benefit from popular parents. However, the nature of how they do this is crucial to reaping the benefits. Maybe it is the similarity in chemical profiles that leads to this effect, in spite of prior related work that might suggest otherwise. In this case, it is in the best interest of cannabis sellers to breed only (or at least primarily) popular strains. Perhaps, however, recognition of relatedness to popular strains through signals in names *does* help this effect manifest itself, but our testing was insufficiently thorough, in which case naming strains accordingly might be the best way to utilize this effect. Either way, only future work will truly let us know.

6.2. Future Work

Note that alternative approaches, further analyses to be made, and related ideas (forms of future work) are all mentioned throughout the paper when relevant. We summarize only the most promising ones here.

First, there are further tests to be carried out to figure out where the parent-to-child popularity retention comes from. It might still be attributable to names, in which case two possible things to try are (1) using not just exact word matches but also semantic similarity using word embeddings, and (2) testing if children strains with "signals" in their names have are closer to their parents in popularity than sibling strains without signals in their names. Alternatively, the popularity retention might be attributable to chemical similarity, in which case studies specifically examining chemical profiles of related strains might yield relevant results.

There are also significant methodology decisions that might be worth re-evaluating. For example, since the two popularity heuristics we used each have evident weaknesses (e.g. average rating being subject to variance due to sample size or review counts not necessarily being representative of popularity but rather recognizability) and often yielded contradicting results, it would be interesting to factor in both of these measures intelligently (perhaps using review counts as a weight for

rating) to obtain a more accurate measure of popularity. It might even be worthwhile to apply standardization methods to these popularity metrics to try regression tests again.

There are also simply other interesting analyses that might be fruitful, such as using named-entity recognition (NER) algorithms to label words with categories and thus perform similar category analyses but with many more categories, or investigating how quickly and in what kind of function the parent-child popularity retention effect diminishes as one looks at grandchildren (of popular strains), great-grandchildren, and so on.

Again, there is little to no work in this field of study, and the apparent senselessness of strain naming has people calling for creating and enforcing strain name regulations or even outright abolishing cannabis strain names altogether [5]. The problem is both neglected and motivated: future work is promising.

Furthermore, with this new dataset available online–the largest, most comprehensive, and most up to date–the problem of understanding strain names is more tractable than ever. Our data is published following the Open Data Institute's standards and practices, and is freely available at this address. We hope this opens the floor to more research in this area.

7. Acknowledgements

First and foremost, thank you to my advisors Professor Christiane Fellbaum and Dr. Jérémie Lumbroso for inviting me to pursue this project, for supervising it, and for their continuous advice and support. Thank you to all students in COS IW 02, as well as our Teaching Assistants Pranay Manocha, Alan Ding, and Gyoonho Kong, for the indispensable feedback and ideas they offered each week. Thank you to Leafly for originally collecting the data used in this project. Lastly, thank you to all my friends who supported me throughout this endeavor.

8. Honor Code

This paper represents my own work in accordance with University regulations.

- Henrique Schechter Vera

References

[1] K. H. Y. S. H. L. T. B. K. J. S. C.-B. A. H. K. Bidwell LC, Ellingson JM, "Association of naturalistic administration of cannabis flower and concentrates with intoxication and impairment," *JAMA Psychiatry*, vol. 77, no. 8, pp. 787–796, 2020.

[2] A. S. F. C. E. T. C. P. Laura Alethia de la Fuente, Federico Zamberlan, "Data from: Over eight hundred cannabis strains characterized by the relationship between their subjective effects, perceptual profiles, and chemical compositions," *Mendeley Data*, no. V1, 2019.

[3] Leafly, "Leafly | about," https://www.leafly.com/news/about, 2021.

[4] Leafly, "Leafly's 100 best cannabis strains of all time," https://www.leafly.com/news/strains-products/top-100-marijuana-strains, 2021.

[5] N. Ponieman, "Cannabis strain names are meaningless: What is the industry doing about it?" *Benzinga*, 2020.

[6] O. C. J. S. H. A. T. H. S. A. S.-H. A. Reimann-Philipp U, Speck M, "Cannabis chemovar nomenclature misrepresents chemical and genetic diversity; survey of variations in chemical profiles and genetic markers in nevada medical cannabis samples," *Cannabis and Cannabinoid Research*, vol. 5, no. 3, pp. 215—-230, 2020.

[7] C. Roberts, "Science reveals the cannabis industry's greatest lie: You're buying weed wrong (and so is everyone else)," *Forbes*, 2020.

[8] Simple English Wikipedia, the free encyclopedia, "List of fruits," https://simple.wikipedia.org/wiki/List_of_fruits, 2021.

[9] A. Wallace, "Cannabis is one industry that's actually coming out of covid even stronger," *CNN Business*, 2021.

[10] Wikipedia, the free encyclopedia, "List of colors by shade," https://en.wikipedia.org/wiki/List_of_colors_by_shade, 2021.

9. Appendix

9.1. Results for Section 5.1: "What strain attributes might help make them popular?"

What follows are a set of tables displaying linear regression model results with independent variables of (1) THC concentration, (2) strain name sentiment score, (3) effect scores, and (4) flavor scores. White rows show results where average rating is used as the dependent variable (i.e. popularity metric), and gray rows show those where review count is used as the dependent variable. Note "R^2" denotes the coefficient of determination for the regression, and "coefficient" and "intercept" denote the resultant linear model itself.

Table 4

Independent variable	R^2	Coefficient	Intercept
THC concentration	0.0019789	0.016063	3.5344
	0.00026688	141.19	-1.7042
Strain name sentiment (compound polarity score)	1.4281e-05	-0.042173	3.6175
	3.837e-05	13.615	67.313

Independent variable	R^2	Coefficient	Intercept
Relaxed (effect)	0.038496	0.071912	4.4704
	0.0060838	-41.786	134.05
Happy (effect)	0.042526	0.076506	4.4714
	0.00023878	8.3795	132.76
Euphoric (effect)	0.022449	0.057599	4.471
	0.0002404	8.7123	132.66
Uplifted (effect)	0.0085529	0.033704	4.4723
	5.359e-05	3.8995	132.86
Hungry (effect)	9.6707e-06	0.0011844	4.4724
	0.0015152	21.669	132.83
Sleepy (effect)	0.00047604	0.0077339	4.4722
	0.00025075	8.2043	132.61
Giggly (effect)	0.0077477	0.032992	4.4716
	0.0003909	10.832	132.59
Creative (effect)	0.0058886	0.028636	4.4729
	9.0745e-05	5.1959	132.97
Focused (effect)	0.0025502	0.018725	4.4729
	0.00079308	-15.263	132.52
Tingly (effect)	0.0057058	0.028629	4.4723
	0.00031485	-9.8299	132.92
Talkative (effect)	0.0028108	0.019956	4.4726
	0.00063567	13.872	133.03
Aroused (effect)	0.0081604	0.036104	4.4735
	7.0449e-05	4.9032	133.03
Energetic (effect)	0.00055528	0.0085631	4.4726
	0.0021682	24.733	133.47

Independent variable	R^2	Coefficient	Intercept
Ammonia (flavor)	1.2323e-05	-0.0014564	4.4724
	0.00049442	-13.484	132.48
Apple (flavor)	0.00044285	0.0076834	4.4725
	0.00051784	-12.144	132.71
Apricot (flavor)	2.7257e-05	-0.0018728	4.4724
	0.0003244	-9.4437	132.78
Berry (flavor)	0.0028148	0.019503	4.4728
	0.00056024	-12.718	132.65
Bluecheese (flavor)	1.7e-05	-0.001566	4.4724
	4.5458e-06	-1.1836	132.87
Blueberry (flavor)	0.00083317	0.01018	4.4725
	7.0569e-06	1.3694	132.88
Butter (flavor)	0.0034111	0.022072	4.4726
	0.00092131	-16.766	132.73
Cheese (flavor)	1.6265e-05	-0.0013119	4.4724
	0.00044452	-10.025	133.05
Chemical (flavor)	3.5802e-06	0.00071982	4.4724
	0.00073777	-15.103	132.61
Chestnut (flavor)	0.00016885	0.0054723	4.4726
	6.656e-05	-5.0219	132.69
Citrus (flavor)	0.0050079	0.024962	4.4722
	0.0026591	-26.587	133.12
Coffee (flavor)	7.2459e-05	-0.0033609	4.4723
	0.00045863	-12.359	132.48
Diesel (flavor)	0.014181	0.042787	4.4727
	0.0019234	-23.032	132.75
Earthy (flavor)	0.0015433	0.014993	4.4731
	0.0011364	-18.805	132.08
Flowery (flavor)	0.0086379	0.037133	4.4741
	0.0027599	-30.68	131.5
Grape (flavor)	0.0013254	0.013156	4.4725
	0.0002574	-8.4741	132.83
Grapefruit (flavor)	2.792e-05	-0.0020085	4.4724
	0.00028752	-9.4211	132.79
Honey (flavor)	0.00044452	0.0082842	4.4726
	0.00051782	-13.069	132.6
Lavender (flavor)	0.00087267	0.01151	4.4726
	0.00060342	-13.99	132.65
Lemon (flavor)	0.0019921	0.015532	4.4724
	0.00080479	-14.43	132.92
Lime (flavor)	0.0068624	0.031881	4.4732
	0.00090869	-16.957	132.47
Mango (flavor)	0.0024308	0.016433	4.4723
	0.00017958	-6.5284	132.91

Independent variable	R^2	Coefficient	Intercept
Menthol (flavor)	2.435e-05	-0.002121	4.4724
	0.00024694	-9.8728	132.68
Mint (flavor)	0.0012013	0.01314	4.4727
	0.00011646	-5.9802	132.75
Nutty (flavor)	0.0051691	0.028626	4.4731
	0.00072943	-15.718	132.49
Orange (flavor)	0.00072341	0.009622	4.4725
	0.0003568	-9.8772	132.8
Peach (flavor)	0.0044251	0.022745	4.4722
	0.00057319	-11.965	132.97
Pear (flavor)	7.8112e-06	0.0013435	4.4725
	0.00016774	-9.1004	132.6
Pepper (flavor)	0.001613	0.01521	4.4725
	0.0022708	-26.379	132.66
Pine (flavor)	0.0024283	0.018261	4.4725
	0.00043004	-11.233	132.83
Pineapple (flavor)	0.0033673	0.020407	4.4725
	2.6482e-07	0.26452	132.88
Plum (flavor)	2.7713e-05	-0.0023201	4.4723
	6.4809e-05	-5.186	132.69
Pungent (flavor)	0.0017856	0.016323	4.4725
	0.00014758	-6.8592	132.85
Rose (flavor)	0.00033606	-0.009067	4.4719
	0.00052342	-16.54	131.99
Sage (flavor)	8.8544e-06	0.0012324	4.4725
	0.00013385	-7.004	132.69
Skunk (flavor)	0.0036137	0.02196	4.4726
	0.00081279	-15.223	132.73
Spicyherbal (flavor)	0.0013691	-0.015065	4.472
	0.00019184	8.2424	133.08
Strawberry (flavor)	0.0022379	0.016527	4.4725
	0.00022084	-7.5885	132.85
Sweet (flavor)	0.0077677	0.033472	4.4727
	0.00013406	-6.4274	132.83
Tar (flavor)	0.0098585	-0.04656	4.4708
	0.000248	-10.794	132.49
Tea (flavor)	0.00064159	-0.0088838	4.4725
	0.00037224	-9.8908	132.96
Tobacco (flavor)	0.0081295	-0.036056	4.4718
	0.00010757	-6.0624	132.77
Treefruit (flavor)	0.0022721	0.019526	4.4732
	0.00092394	-18.2	132.15
Tropical (flavor)	0.0064829	0.029813	4.4727
	0.00068752	-14.191	132.75

Independent variable	R^2	Coefficient	Intercept
Vanilla (flavor)	0.0053821	0.025227	4.4721
	0.0010956	-16.637	133.08
Violet (flavor)	0.0011898	0.015233	4.473
	0.00030356	-11.247	132.47
Woody (flavor)	0.00015766	-0.0048562	4.4723
	0.00072448	-15.216	132.37