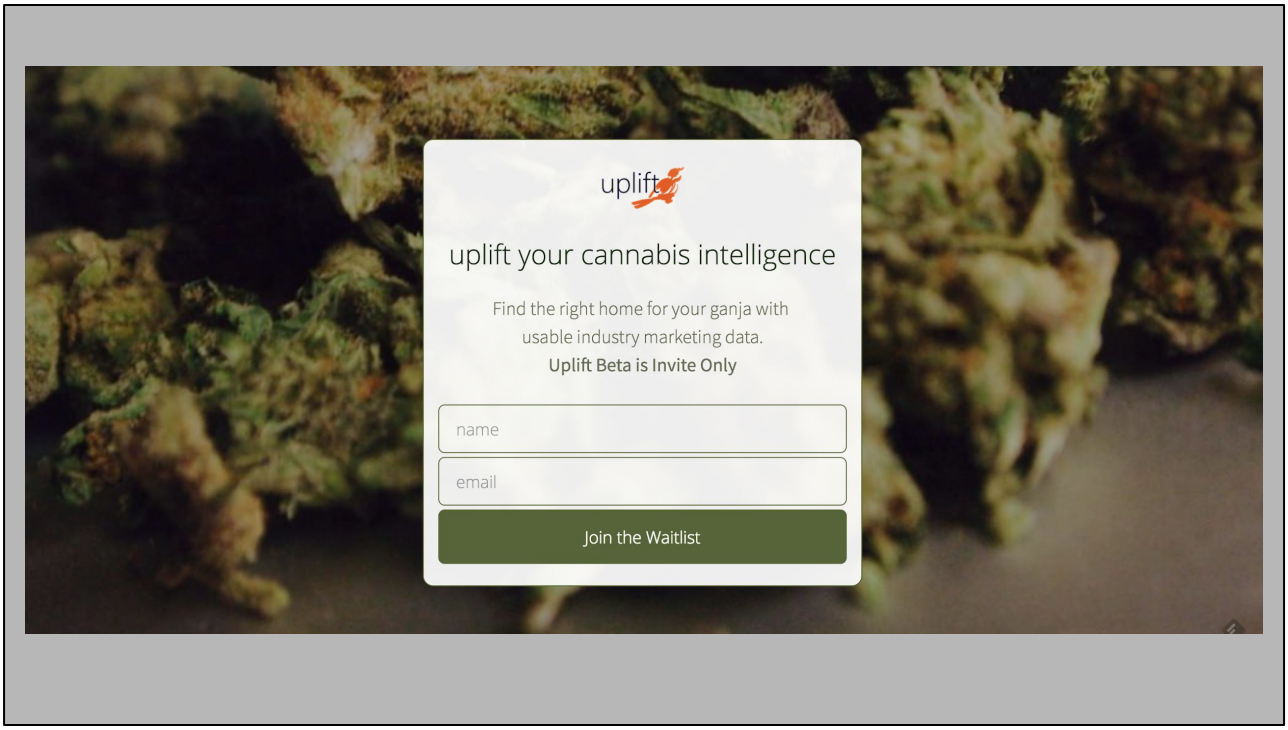
Picking Weeds:

A Looking Glass into the Undiscovered Sales Trends in WA Cannabis

Steve Wald, PhD
Bluefish Analytics, LLC
steve.wald@bluefishanalytics.com
linkedin.com/in/steve-wald/

In this presentation, I will be demonstrating EDA and vizualization tools I'm developing for Uplift BI, LLC, a business-intelligence platform co-created by Max Caughron, PhD, for clients in the legal cannabis business. The tools I'm demonstrating this evening provide a looking glass into differential sales trends among the kaleidoscopic array of products circulating in the industry.



Business Context: The legalization of cannabis for recreational use in Washington state and other states has transformed what was once one of most opaque of industries--an illicit black market--into one of the most transparent. How did this come about? The transformation arose because of the detante that exists between federal law enforcement, which regards cannabis as a controlled substance, and states with legalization such as Washington. This detante, encoded in the famous “Cole Memo” of 2013, rests on the commitment to maintain an iron wall between business activities regarded as legal by the states and activities that everyone regards as illegal--such as the association of the trade in marijuana with violent crime. In Washington state, this iron wall is maintained through rigorous and exacting TRACEABILITY PROVISIONS imposed on all players in the legal trade.

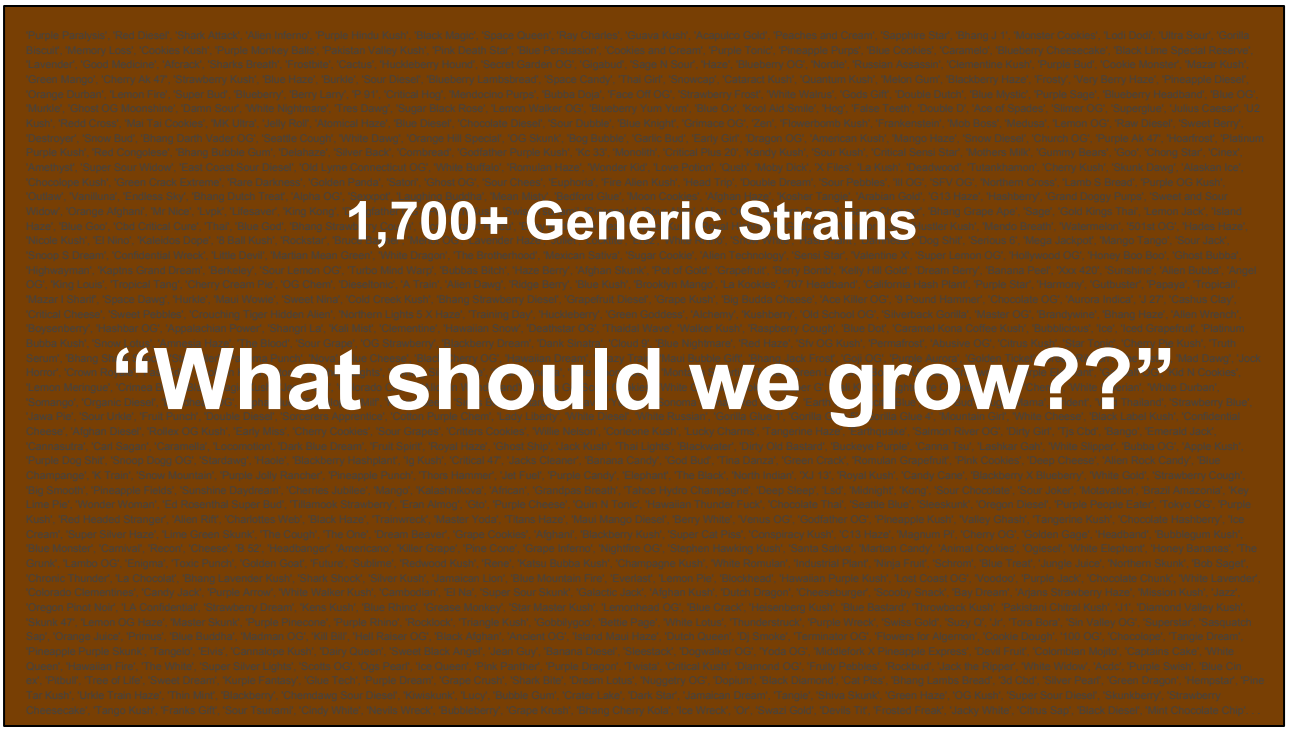
These provisions require producers, processors, laboratories and retailer/dispensaries to track each and every “transfer” of a cannabis product from seed to sale through daily and weekly reporting to the state. For example, buy something in a pot shop, and a record is created that carries with it a host of information associated with the exact origins, quality and pricing of the cannabis contained in the product purchased.

By virtue of the state’s reporting requirement, this great wealth of economic data is entered continuously in the PUBLIC RECORD. But just because it’s public does not mean that it’s ACCESSIBLE. Quite to the contrary. Accessing this data is a protracted process that begins with a FOIA request to the state, followed by delivery of files containing massive, ambiguous and messy datasets comprising tens of millions of

records. The complex and intricate pipeline necessary to wrangle this data and the development of a proprietary relational database to order it for analysis constitute a significant part of the intellectual property developed by my client, Uplift BI, LLC.

Once this data enters the proprietary database, we want to bring our business problems to it. Here’s a salient one: Walk into a dispensary, and one feature of the industry that will strike you is the incredible variety of colorfully branded strains of cannabis available to purchase. In it’s variety, cannabis gives heirloom tomatoes a run for their money…

[Click for multiple animations and to advance to next slide.]

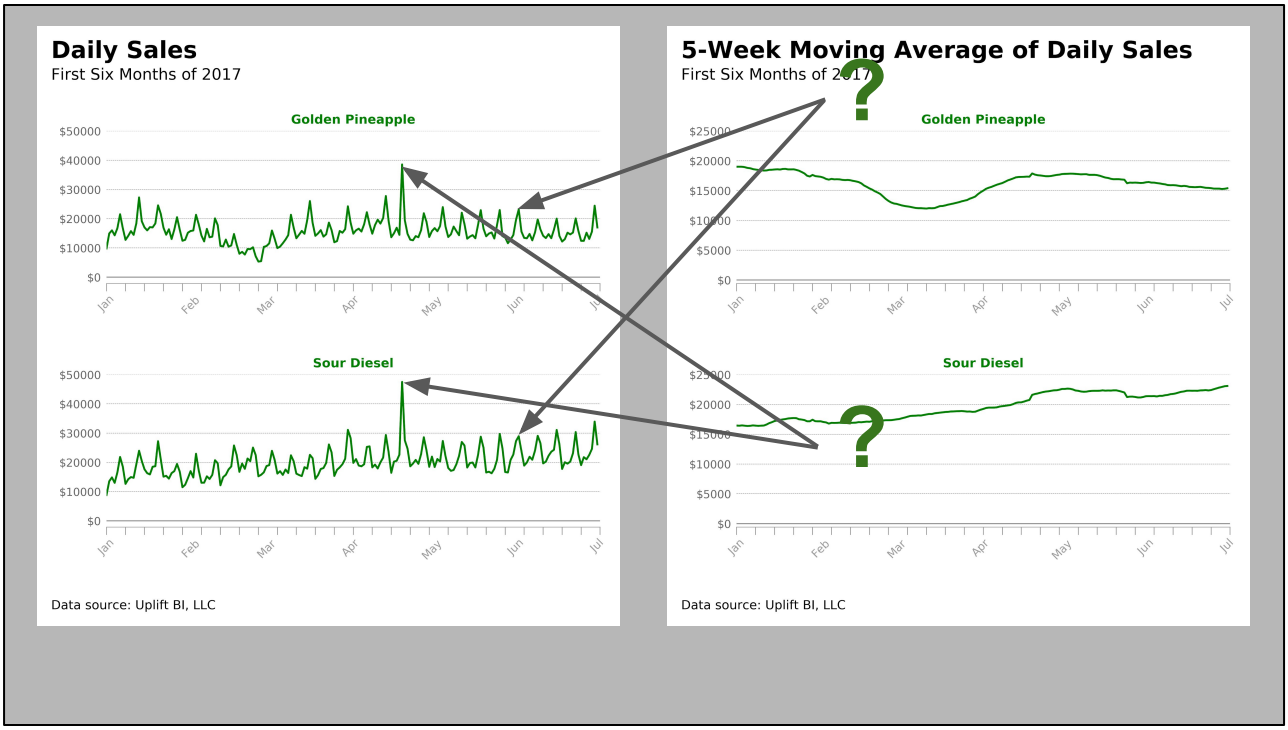


All told, there are over 1,700 registered generic strains of cannabis. If I’m a grower, having the right product hit the market at the right time is key to maintaining my competitive advantage. This all begs the question: [click for animation] What should we grow? This is a question that can be informed by the observed performance of various strains on the market. Using trends in aggregated daily sales (or units sold) as a proxy for performance, we can delve into Uplift BI’s database for insights into what products are trending (or not) in the market.

Product Specifications: I was tasked with developing a set of data containers and dynamic visualization tools that would allow consultants at Uplift BI to easily and quickly explore sales trends among products within the database. The viz tools would need to provide flexibility necessary for production-quality graphics in static reports delivered to Uplift BI’s clients--primarily producers and processors in the industry. I was also tasked with defining meaningful trend metrics by which products could be automatically ranked.

The demo and discussion that follows will review four such performance metrics and the methodology underlying them. The database used for demonstration purposes covers state-wide sales statistics for 21 strains, with data from 2015 to mid-2017. For this demo, we will often compare strains in pairs, because theoretically, if we can rank two strains, a machine can easily rank 2,000.

So let’s look at some data . . .



First two strains in the ring: Golden Pineapple and Sour Diesel. Two features common to the vast majority of trends in the database:

[Click for animation 1] Day-of-week (DOW) seasonality--surges on the weekends, and . . .

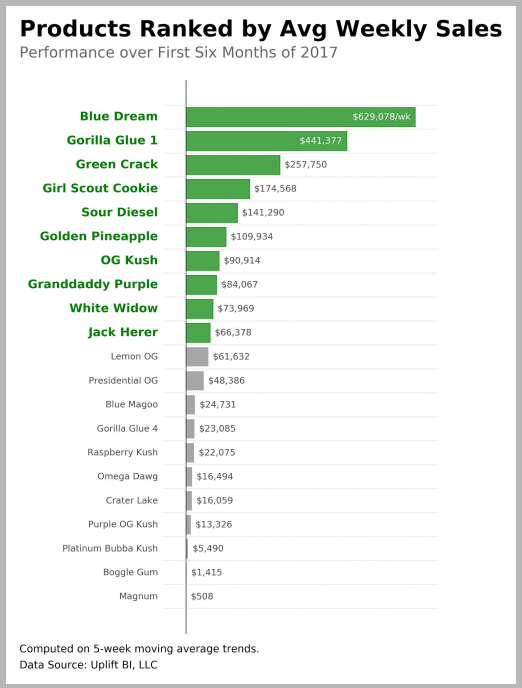
[Click for animation 2] The annual 4-20 weed holiday. DOW seasonality is noise; hence in the industry, WEEKLY SALES (versus daily) is regarded as the elemental unit of performance. Since the 4-20 bumps are common to most strains, we can keep this in the data.

[Click for animation 3] A solution to dealing with DOW noise and the 4-20 bumps is to smooth the trends with a moving (rolling) average. Trends on the right show a 5-week moving average. Choice of the 5-week “boxcar window” versus some other length is largely a matter of statistical art, not science. And the tools provide flexibility to adjust the window. You want a window wide enough to smooth out volatility and noise but small enough to register (in the business case at hand) weekly movements in trends.

From the 5-week MA trends, we can eyeball it that Sour Diesel had a better period, but we still need a quantitative measure to rank. Thus, our first metric . . .

Metric 1

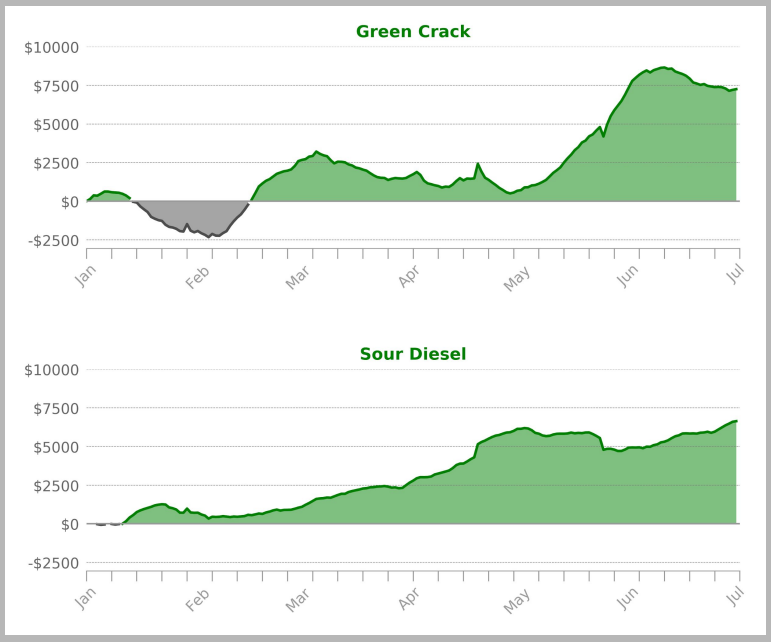
“Avg. Sales”



Divide the sum of sales over the period by the number of weeks in the period. This gives us a basic metric of average weekly sales, which has value in providing a conspicuous measure by which to cluster different products into classes.

We see that Blue Dream is the “Big Mac” of the weed world; the inaptly named “Magnum” at the bottom . . . not so much. That may be because Magnum is unpopular, or because it’s a boutique strain only available in limited times and places. We don’t know, but we can see that these two products fall into different classes.

Big problem! This metric does NOT convey any information about DIRECTIONALITY. Blue Dream may be selling a lot, but is it gaining or losing ground over the period?



One way to put this directionality into relief is to return to the trend lines and shift them to a baseline. One appropriate baseline would be the median value of sales for each product over the sample period. However, because we are using a moving average, we can also set it up as a race by shifting the trends to zero at time-zero and seeing where each goes from there. These graphs of trends shifted to baseline make it even more evident that Sour Diesel had a better period than Golden Pineapple.

[Click for animation] But what about this pairing? Not so easy right off the bat to declare a winner between Sour Diesel and Green Crack. How can we develop a quantitative measure to compare these two squiggly lines?? (Next metric)

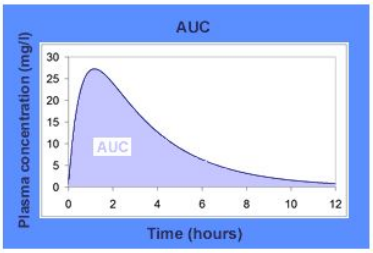
Metric 2

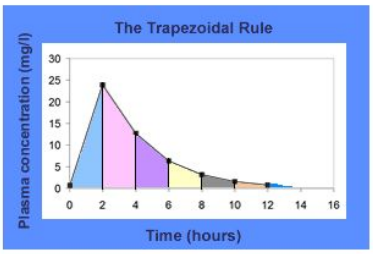
“Uniform Gain”

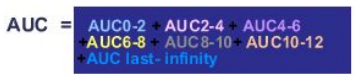
Trend Metric: Area Under Curve (AUC)

AUC ~ total bodily exposure to drug after dose
unit: mg(drug) x hour / L(plasma)

```
np.trapz(series)
```



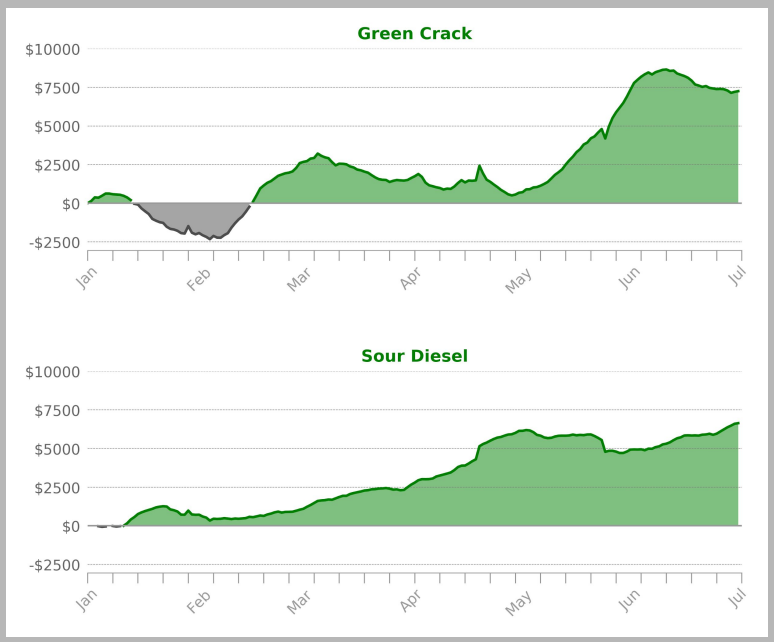




Source: Univ. de Lausanne, “Area under the Plasma Drug Concentration Time Curve,” https://sepia.unil.ch/pharmacology/index.php?id=66, accessed Feb 26, 2018

We’ll do so by borrowing a method that pharmacologists use to measure the diffusion of a drug in the body over time after dosage. In studies of drug metabolism, a drug is administered, and then its concentration in the blood is measured at even (e.g., hourly) intervals. These data produce a curve, and the AREA UNDER CURVE (AUC) represents the total exposure of the body to the drug over time. This measure is in a weird unit: “milligram-hours” per Liter of plasma (to be discussed soon).

The trapezoidal rule (bottom graph) is one method commonly used to estimate AUC. Simply drop vertical lines from each datum to the x-axis and add up the area of all the resultant trapezoids. [Click for animation] The NumPy module in Python provides a super-easy method to compute this value from a Pandas time series.



AUC

≈ 422,848 ?

dollar days??

≈ 569,704 ?

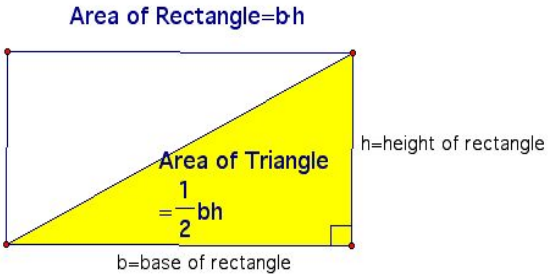
Applied to our contestant strains [click animation], we can now see that Sour Diesel once again wins! It has approximatly 147,000 more AUC . . . but 147,000 of [click animation] what? [click animation] “Dollar-days”? What does that mean? Notice, NOT “dollars per day.” That would be downright intelligible! Nope, “dollar-days.”

It turns out that through a simple transformation using middle-school geometry, we can transform the AUC metric into something meaningful . . .

Imagine the total area under each curve as a specific quantity of liquid--say, green smoothie. We pour each quantity into its own vessel, and then we're going to pour each of those under straight lines anchored at the origin of the graph: T-zero = 0. The more fluid poured under the line (or OVER IT, if we're dealing with negative AUC), the steeper the line will tilt. Steeper is better (again, if we're talking about POSITIVE AUC). We want, then, to compare the slopes of those lines.

Deriving Uniform Weekly Growth Rate from AUC

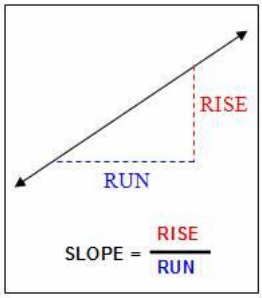
Area of a right triangle



$base(days)$

Slope of a straight line

Use "height"($) over "base"(days) to find rate.

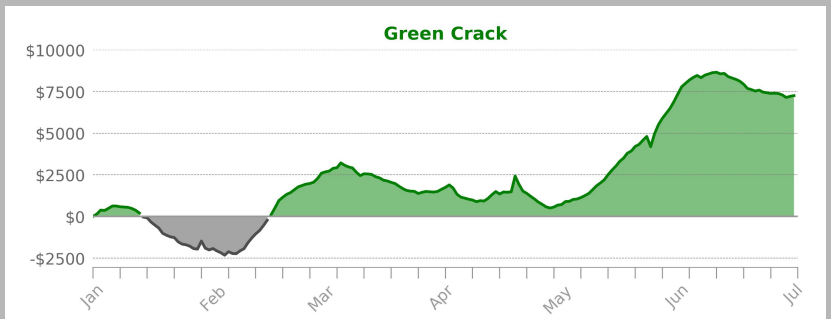


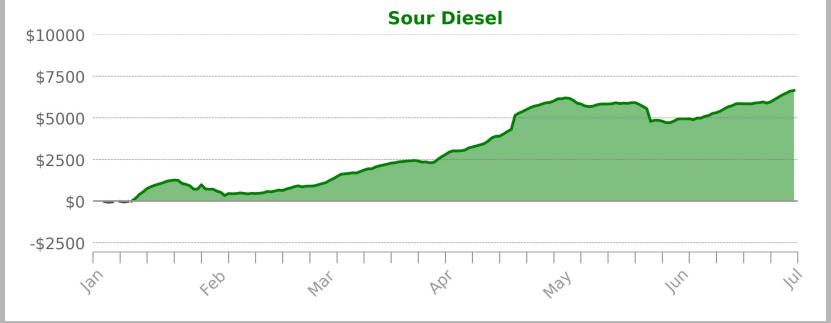
$$= X\frac{\$}{day}$$

$$= 7X\frac{\$}{week}$$

When we're done pouring the smoothie under a line, our new graph will look something like the diagram on the left, the formula for the area of a right triange. We want to know the slope. We know the base (a certain number of days) but we don't know the height. HOWEVER, we DO know the AREA; it's the AUC. [Click animation] So we simply rearrange the formula to solve for unknown height and substitute the value for AUC for area (A) in the formula. Keeping track of our units, we see that the 'days' units cancel out on the right side of the equation, leaving us a height in dollars.

[Click animation] Then we simply plug that height into the slope formula as the 'rise' over the base in days as the 'run.' We end up with X dollars per day or 7X dollars per week! This figure represents the irregular gains or losses of the squiggly trend redistributed throughout the period as a "uniform gain."



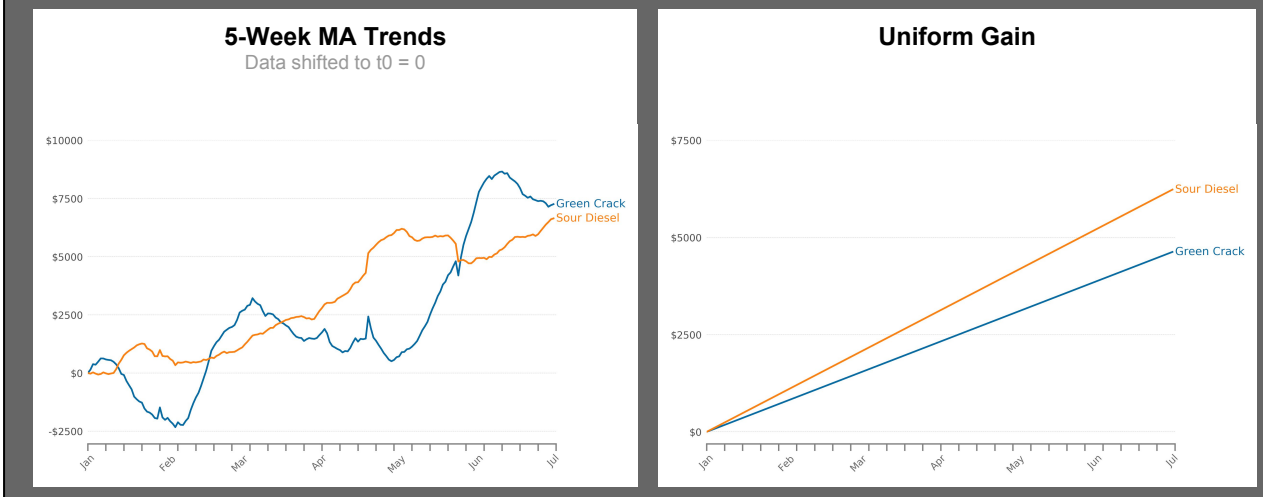


Uniform Gain(Loss)

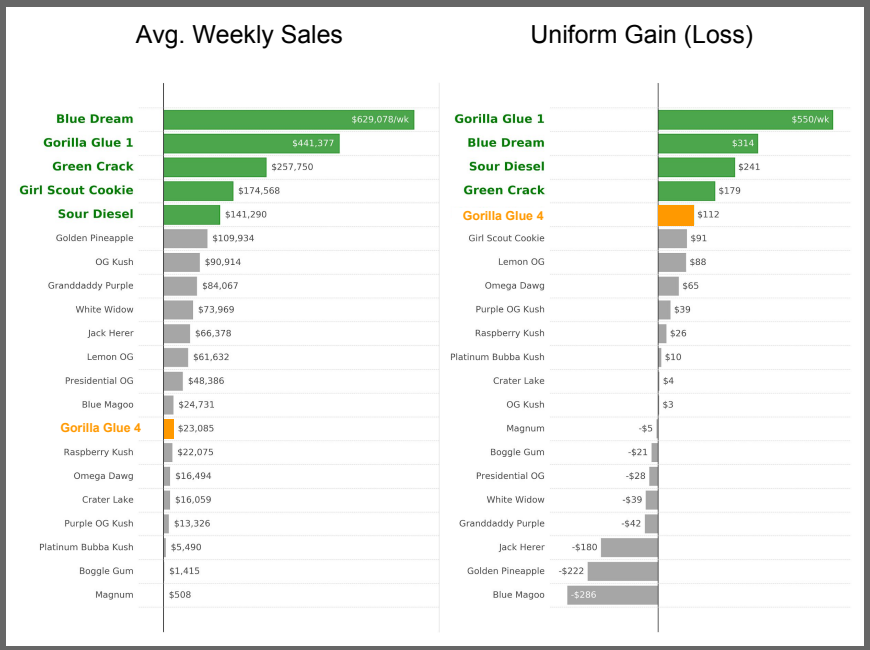
≈ $179/wk

≈ $241/wk

Here's what this transformation looks like for our two contestants [click animation] . . .



Here's another look: On the left, hard to rank by eye; on the right [Click animation], easy!



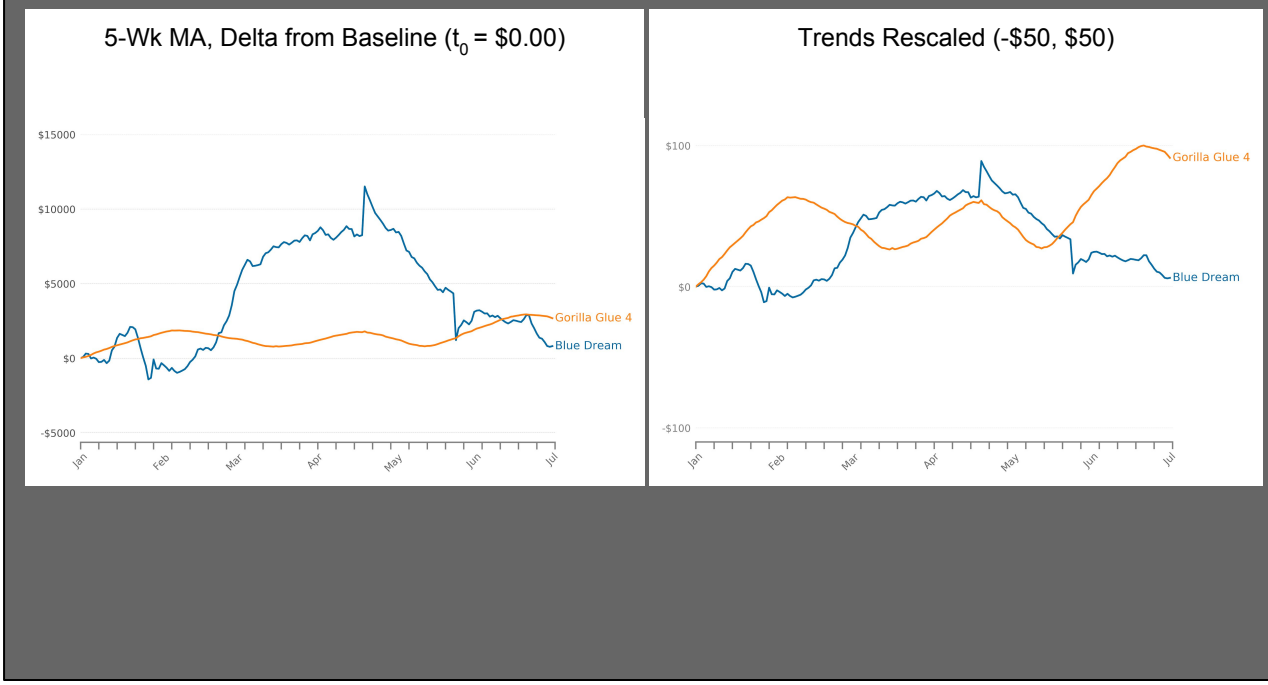
Here are our 21 strains ranked on the left, for reference, by sales, and listed in the same order on the right by uniform gain. We can see that the rankings over these two metrics diverge. [Click animation] Here they are ranked for both metrics.

[Click animation] We discover that a lower-volume strain, Gorilla Glue 4 (ranked 14th in sales) is keeping pace with the big boys in terms of the momentum its gaining over the period in increased sales. This is in actual dollars to scale.

We might want to compare a strain such as Gorilla Glue 4 to a large-volume strain in such as way as to offset the difference between the two in their sales overall volumes that place them in distinct and separate classes. This is much like in the stock world where we compare the price of two stocks not by their absolute prices (which are arbitrary) but by the ratio of price to earnings. This brings us to our third metric . . .

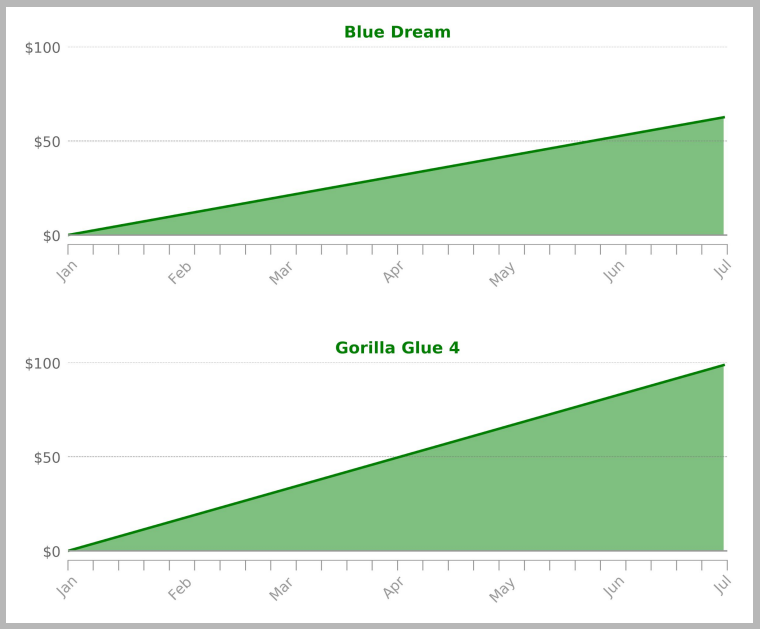
Metric 3

“Relative Rate”



One way to offset for volume is to resize each trend to a uniform scale. Let's say for each trend, we call its worst day -$50 and its best day +$50. The scale is arbitrary, a matter of convenience. Then we once again shift both rescaled trends to the baseline. Here we have the UNSCALED lines of the "Big Mac" strain Blue Dream against our little performer, Gorilla Glue 4. This is clearly an apples-oranges comparison; the relatively high volume of Blue Dream means that even small percentile shifts in the trend result in relatively large amplitude changes compared to its smaller rival strain.

[Click animation] On the right, we see the rescaled trends. The effect to rescaling has been to exaggerate the movements of the smaller trend and tone down those of the larger one.

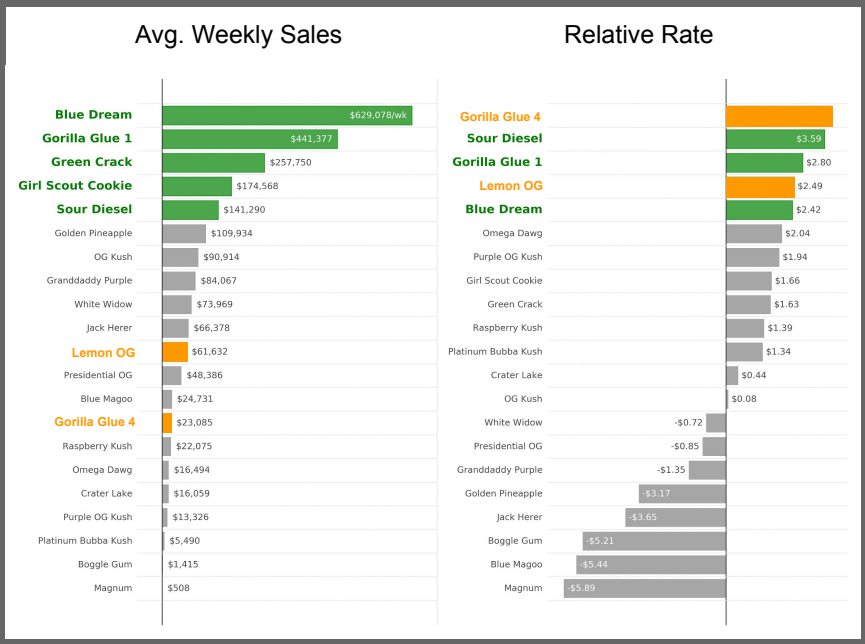


Relative Rate (from AUC)

≈ $2.42/wk

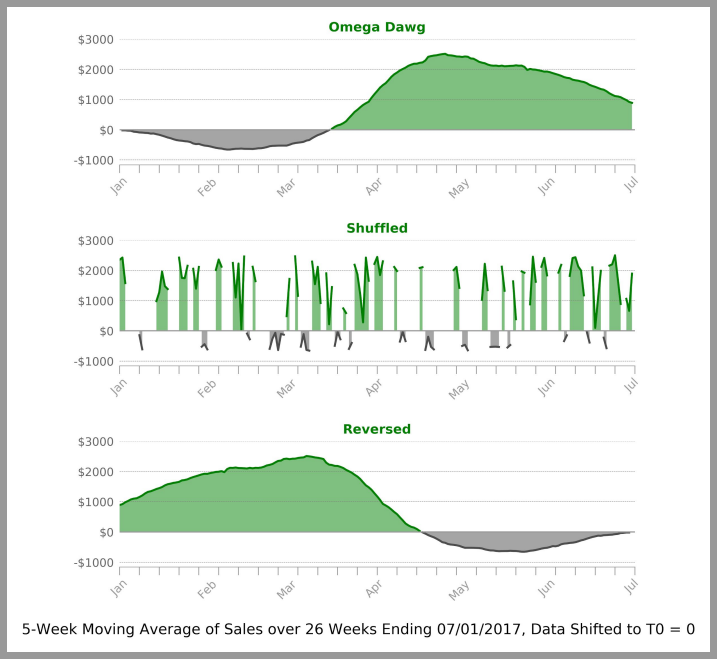
≈ $3.81/wk

Here's another look at these rescaled trends. And here [click animation] is a look at them after using our AUC trick. We can see that, FOR ITS SIZE, Gorilla Glue 4 outperforms Blue Dream over the period. The units, in dollars, are simply a result of the scale we chose--you could call it uniform gain in "Monopoly money."



Here are our 21 strains ranked by relative rate, again showing the divergence of rankings among the two metrics. [Click animation] And here in rank order by each metric. [Click animation] We now see our ‘hero’ Gorilla Glue 4 at the top of the heap. And we make another discovery: [Click animation] Lemon OG is also gaining ground at an impressive rate for its modest size in overall sales volume.

AUC Caveat!



AUC

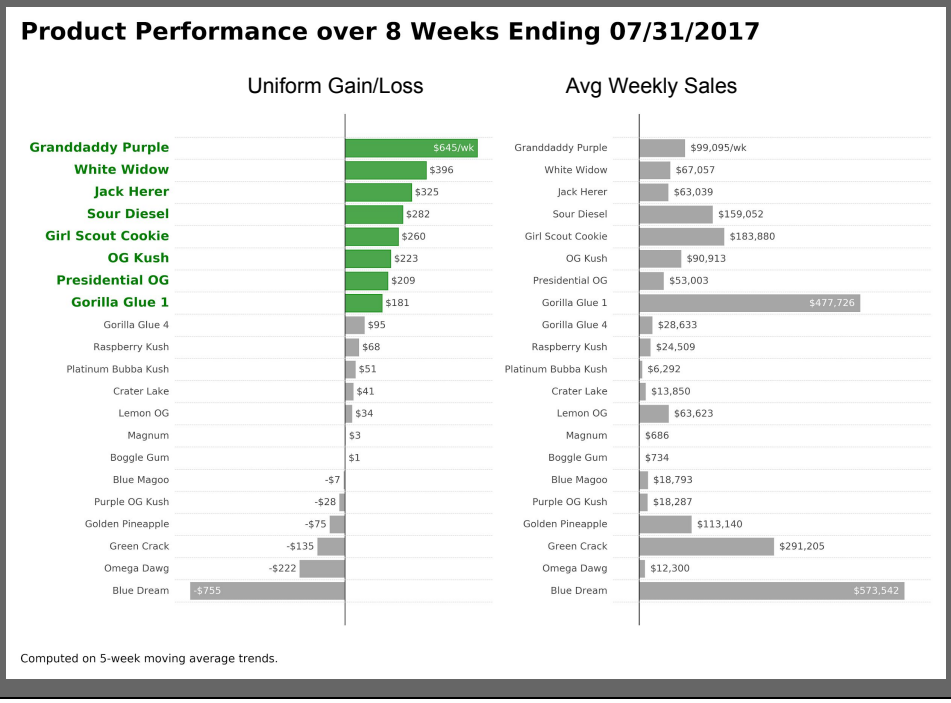
≈ 152,658 !

≈ 150,982

≈ 152,658 !

One problem with the AUC method illustrated here: The SEQUENCE of data MATTERS for the AUC. [Two click animations] And in fact the AUC metric penalizes trends for higher volatility. BUT . . . THE ORDER does NOT MATTER. [Click animation] The AUC of Omega Dawg is exactly the same as that of the very same trend [Click animation] IN REVERSE! And in the case of the bottom graph, we might like to know that a well-performing strain experienced most of its good performance only at the BEGINNING of the sample period. Yikes! More recent data would suggest that “Reverse Omega Dawg” might not be such a great bet for a grower.

How do we mitigate this potential ambiguity in assessing the strain? Of course, we’re going to look at the actual trend lines before placing a bet on the next 10-week growing cycle. But there are other ways a machine can do so . . .

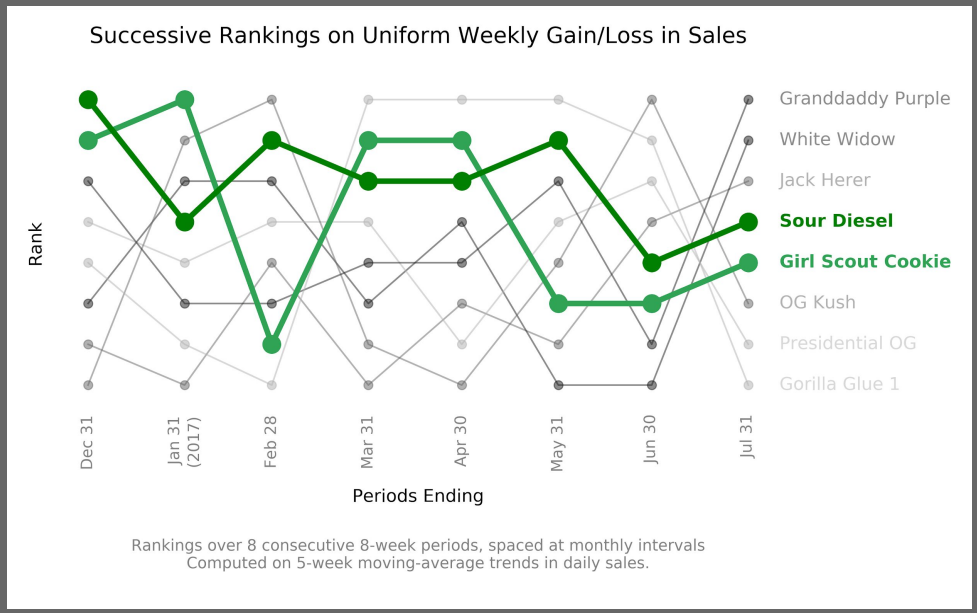


One way is to shrink the aperture of the sample period. We might be more interested on the most recent SIX WEEKS performance than the most recent SIX MONTHS. Let's use a hypothetical use case to illustrate. I'm a grower and I want to identify and select among the top eight strains in terms of absolute growth (i.e., I don't care about differences in overall volume per se; I want strains that are adding the most dollars over a period.) So I'll use the "Uniform Gain" measure as opposed to "Relative Rate."

On the left are the fastest growing trends in real dollars for the most recent eight weeks. Granddaddy Purple comes out on top.

Metric 4

“Back-tested Bestsellers”



But Grandaddy Purple won out only for the most recent eight-week period. What if we had run the same test a month earlier, and another one a month earlier than that, and so on several times? The strains that most consistently rise to the top--like “weeks on the Bestseller List”--would seem to be better bets. This first image shows the ranking performance of our top-eight gainers over several consecutive 8-week periods spaced one month apart. (First image.) Very hard in this image to see which had the best performance over the period; there was quite a bit of shuffling among rankings from one test period to the next.

But if we simply add the total rankings for each product over each of the test periods, the strain with the smallest sum shows the most consistent strong performance. We can see that here. [Click animation] Sour Diesel, once again, is the champion, with Girl Scout Cookie not far behind.

(The End)

Q&A

Big thanks to . . .

Max Caughron, PhD
Uplift BI, LLC

Alex Brooks
Entreprov, LLC

Steve Wald, PhD
Bluefish Analytics, LLC
steve.wald@bluefishanalytics.com
linkedin.com/in/steve-wald/

The tools just demonstrated in their ‘base model’ are generally applicable to any set of sales trends or even to other time series such as trends in supply. Another important feature of the toolset beyond the ‘base model’ presented here is the ability to filter data by geography and specific businesses such that one could easily see, for instance, which “kush” strains trended best during the summer of 2017 in, say, Tacoma. Or, which retailers have shown the strongest overall sales growth in Snohomish County over the recent quarter.

Tech stack: Postgres > Python (NumPy, Pandas) > Matplotlib

Thank you! Questions?