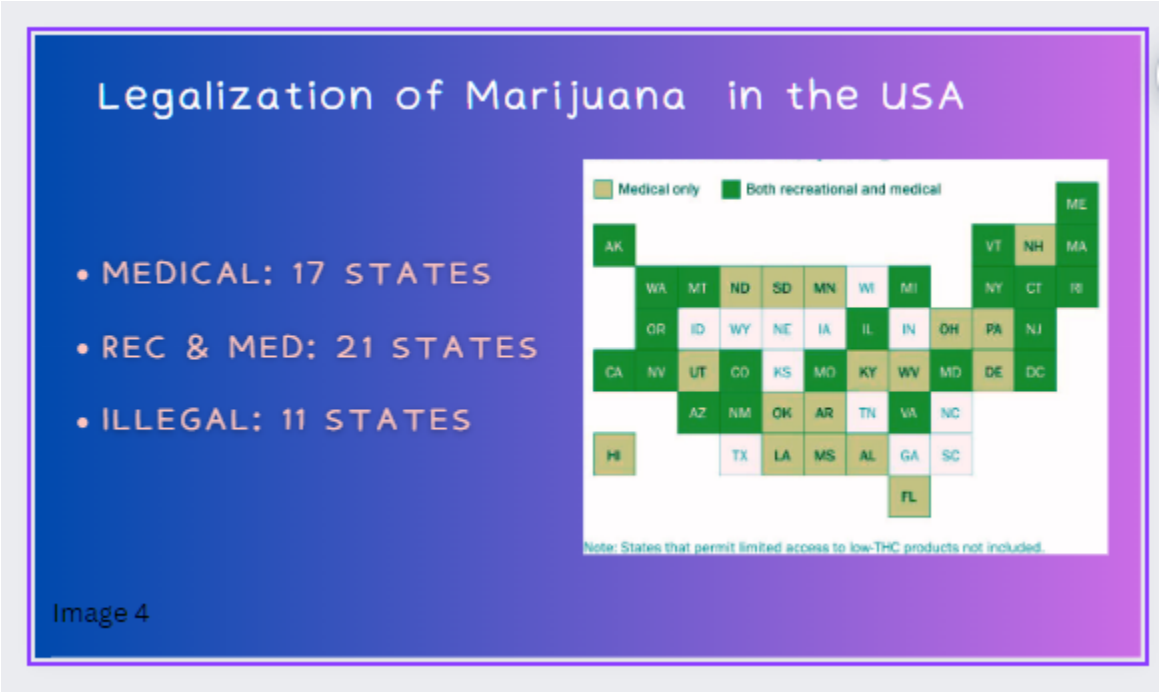
SMU Data Science - Professor Booth
Project 1 - Group 2: Fall 2023

Group Members: Misha Borunda, Kimberly Childers, Lattiana Escobar, Shanara Hawkins, and Tacha Robbins

Do You Love Me, Mary Jane? Insights Uncovered from Patterns and Trends Found within the world of Legal Cannabis

After more than 2500 years since its first ever documented use case in the mountainous regions of Asia, a growing body of current, global research is proving that the popularity of cannabis is steadily climbing towards the top. Society's attitudes surrounding its legalization for medicinal and recreational use is shifting towards a favorable reception, particularly in the United States. For instance, the visual below titled, "*Legalization of Marijuana in the* USA," indicates that as of April 2023, more than 75% of the country has established laws legalizing the use of cannibus for medical or recreational purposes, or both. The cannabis landscape is changing rapidly, which is the source of inspiration driving our team to press further, seeking potential insights and trends influencing its widespread popularity and shifting perspectives.



In light of the expanding legislation and debates where cannabis is at the forefront, our analysis pinpointed several critical research questions highlighting the multi-faceted ways where cannabis could be impactful. For example, what societal, economic, or health implications are at play when states decide to resist the legalization of cannabis? How does the dynamic between breeders and their consumers change or evolve in states where cannabis is

recreationally and/or medicinally legal vs. states where it is currently in dileberation? The goal of addressing these questions is to understand the various government, industry, and societal outlooks that has made cannabis use more appealing in our contemporary society.

To ensure a cohesive and comprehensive understanding of cannabis in the context as it is presented throughout our analysis, it is important to define the following key terms addressed in our data set: types, indica, sativa, hybrid, effects, ratings, flavors, strains and breeders. The chart below includes the key terms in the left-hand column and its corresponding definition in the right-hand column. Clarity of the terminology will serve to define the concepts as they are raised in the dataset as well as each of the four hypotheses that follow.

KEY TERM	DEFINITION
TYPES	Cannabis is categorized into three main types: indica, sativa, and hybrid
INDICA	A typy of cannabis that typically produces a calming effect
SATIVA	A type of cannabis that typically produces cerebral effects like creativity & focus
HYBRID	A mix between indica and sativa. Mixes can vary in ratio (50/50, 70/30, etc.)
EFFECTS	A feeling or mood followed during or soon after cannabis consumption
RATINGS	A numerical assessment of the perceived effect/flavor associated with a strain or type of cannabis. Ratings are on a scale of 1 - 5. One is the lowest & five is the highest
FLAVORS	A particular taste used to alter or enhance the cannabis' flavor profile
STRAINS	A unique combination of compounds found within the cannabis (THC, CBD, etc.) All strains are categorized by one of the three types.
BREEDERS	A business that or a person who cultivates or develops strains of cannabis plants

Our original dataset is titled, "Cannabis Strains," which we pulled from the Kaggle.com website. The contents contained within this particular dataset included the strain, type, rating, effects, flavor, taste, and descriptions. Our main goal during the data cleaning process was to preserve enough information to maximize the validity and reliability of support for our findings. We refined our columns and extracted breeder information from the "description" columns resulting in a robust amount of data, comprised of 2351 rows, used to formulate our research questions, and examine and interpret the validity of our hypotheses.

Using the cleaned and final dataset as a guide, a total of four hypotheses were generated, all serving as focal points helping to unravel the informational diversity and key cannabis insights discoverd through data analysis. Following are the four selected hypotheses for continued investigation: [1]*Do hybrid types have more happy effects compared to non-hybrid types?* [2]*Does the sweet flavor cannabis have the highest ratings?* [3]*Do the breeders with the highest count*

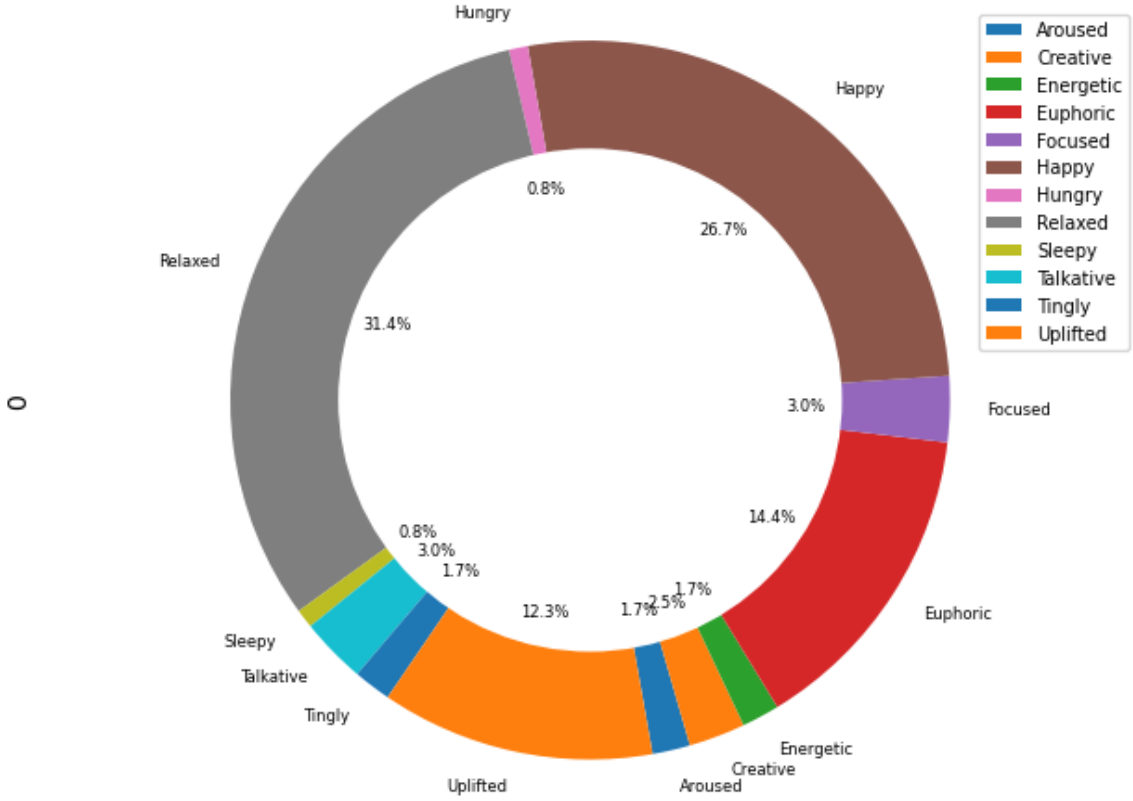
of unique strains have the highest average ratings? [4]*Is there a significant difference in ratings by type or ratings by effects*? With these hypotheses as our guide, our thesis seeks to analyze the ways in which a cannabis-related dataset, encompassing information on a diverse set of factors such as type, flavors, mood, effects, strains, breeders, and user ratings, emphasizes the comparative insights related to the legal production and consumption of cannabis.

HYPOTHESES 1-4 DISCUSSION & ANALYSIS

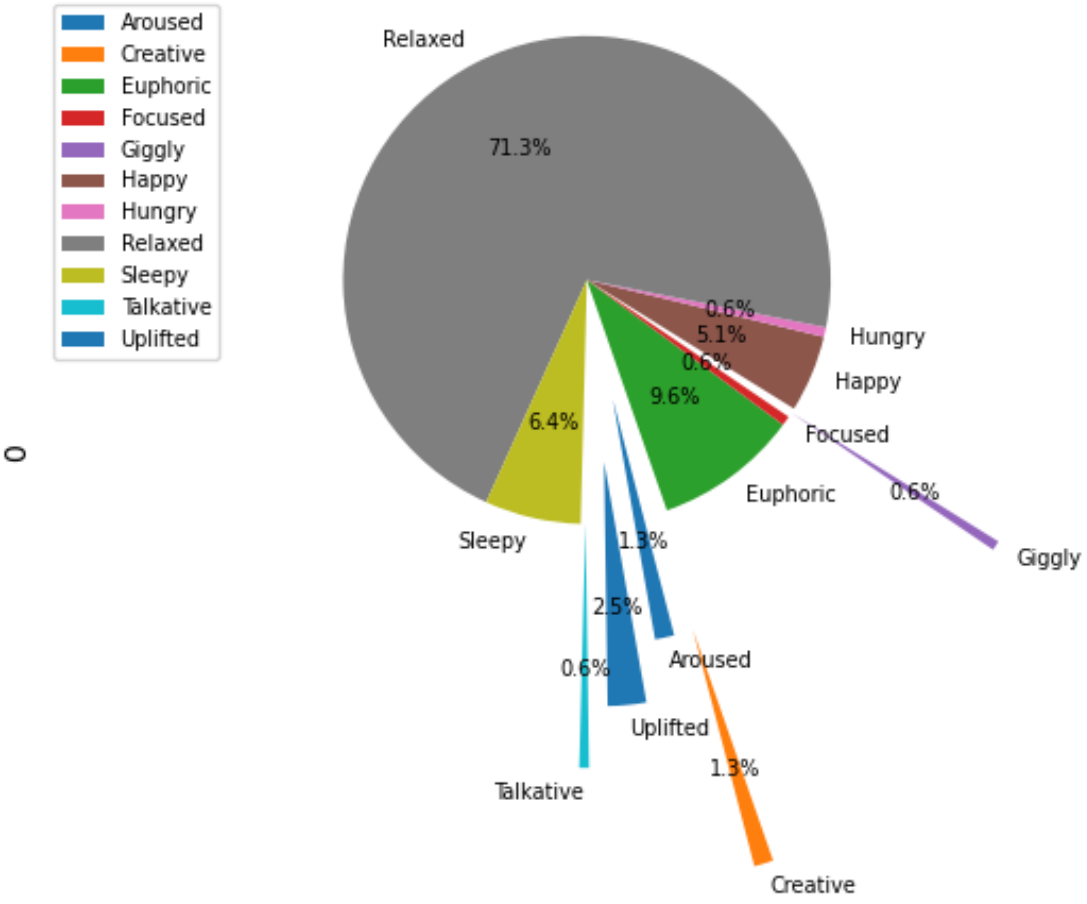
Hypothesis 1: Discussion and Analysis

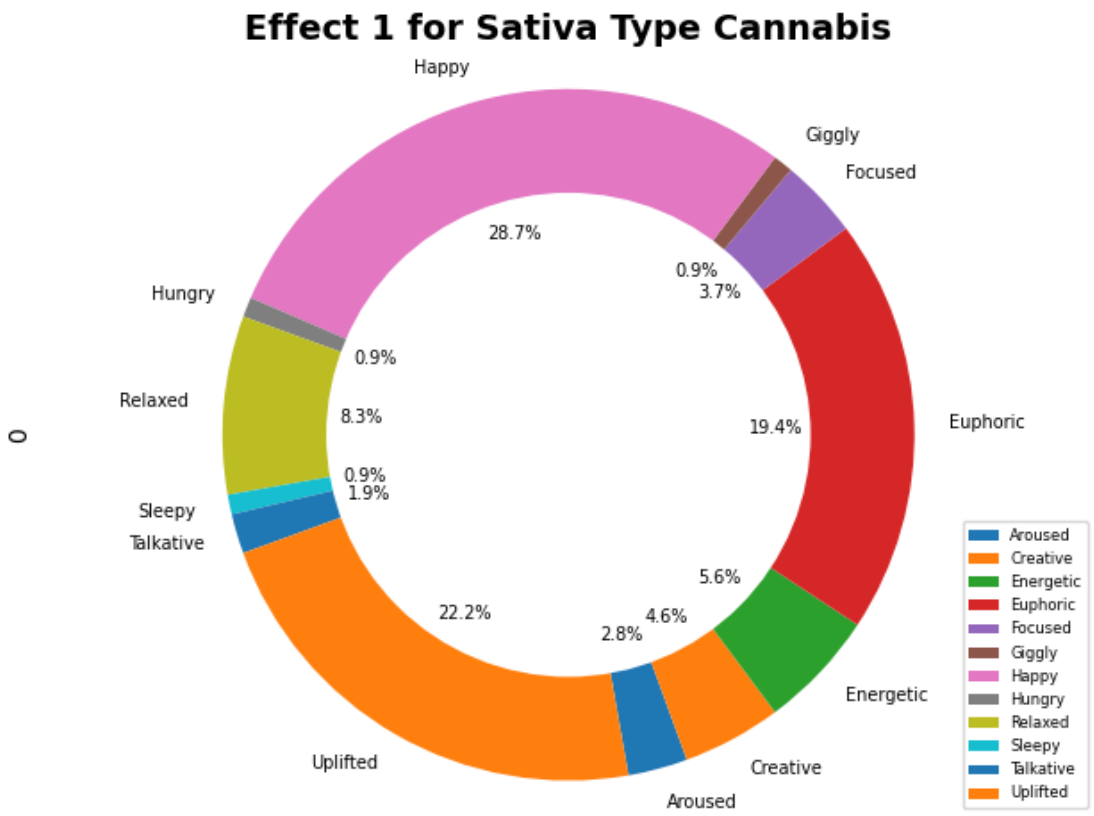
Each cannabis type lists numerous effects per cell within a data row. However, the following analysis of hypothesis number one is solely focused on the initial effects reported listed from first to last within a given cell. Next, to identify the main effect experienced by each type, it was necessary to get the total count of values for each effect. The extracted data values were subsequently plotted on two types of data visulations including both donut and pie charts. each reflecting the total percentages of feelings reported as it specifically relates to the type of cannabis. This information indicated that the main effect experienced by hybrid cannabis type was reported by consumers as “relaxed” with 31.4% of the values, and “happiness” experienced the second most common reported effect with a value of 26.7%. Therefore, this hypothesis proved to be incorrect. In fact, the other type of cannabis, Indica, also highlighted “relaxed” as the main effect experienced with a higher percentage of 71.3%. One noteworthy factor of interest to pinpoint is that the type of cannabis indicative of “happiness” as its main effect is Sativa with a reported value of 28.7%.

Effect 1 for Hybrid Type Cannabis

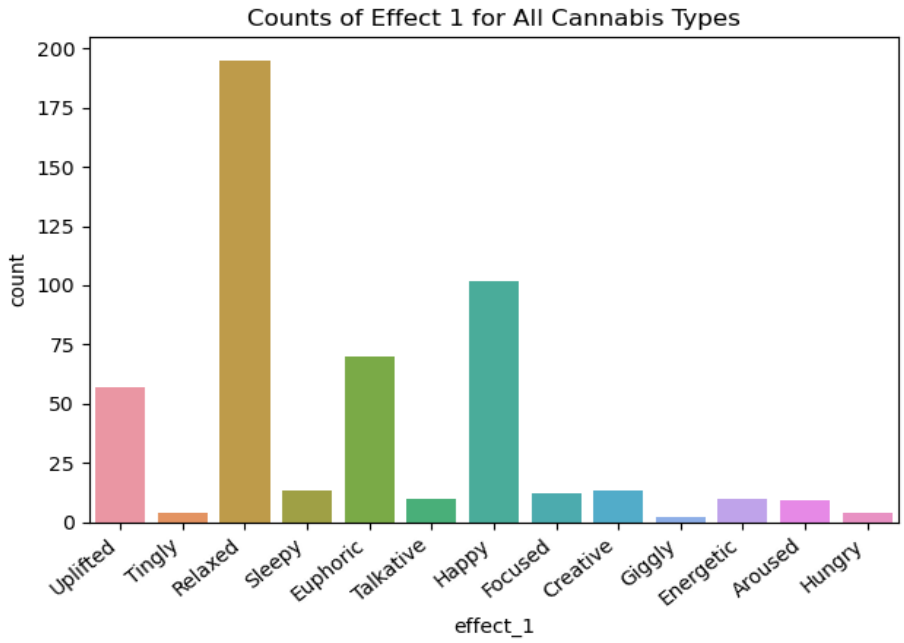


Effect 1 for Indica Type Cannabis





While working on hypothesis one, it was interesting to see which of the main effects would be reported as the first listed effect when consuming Cannabis regardless of the type. To figure this out, we decided to take the total count of reported effects and compare the results. To visually display this information, we created a simple, yet easy to read bar graph, which can be viewe below. The results showed that “relaxed” was the main effect experienced, followed by “happy” and in third place was the “euphoric” effect. Surprising, and of interest to note, “giggly” was the least reported effect experienced by consumers. This is a finding of particular interest given that stereotypically “giggly” is an effect most commonly reported by cannabis users anectdotally in previous the decades prior to the legailization of cannibus consumption recently established in the United States.



Hypothesis 2: Discussion and Analysis

There are a total of four flavors that were used. Team 2 compared these four flavors to their respective rating. After cleaning the data, Team 2 realized that flavor four only had 43 data points out of 2351 entries. The team decided to drop this flavor four dataset due to it being so small compared to the other flavors.

The second hypothesis the team posed was that the sweet flavor cannabis would have the highest ratings. Typically, when looking at flavors one would think that the population would lean towards favoring sweet flavored products as seen with popular vape flavors. Due to this assumption, the team decided that sweet flavor cannabis would be the most popular flavor and have the highest rating.

The raw data was cleaned to break out the flavors into flavor one, two, and three by taking the first listed flavor in the entire “Flavor” column listed in the original data frame (reference figure below). Those flavors made up the flavor one profile. The same steps were repeated to make the flavor two and three profiles. This data is what was used for statistical analysis.

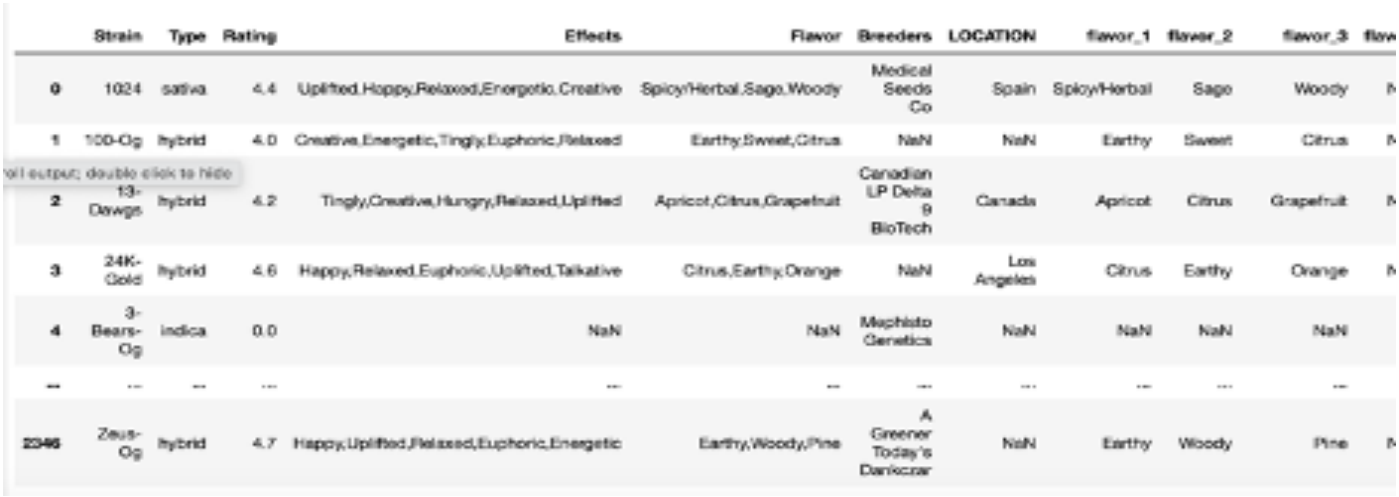
Out[4]:

	Strain	Type	Rating	Effects	Flavor	Breeders	LOCATION
0	1024	sativa	4.4	Uplifted,Happy,Relaxed,Energetic,Creative	Spicy/Herbal,Sage,Woody	Medical Seeds Co	Spain
1	100-Og	hybrid	4.0	Creative,Energetic,Tingly,Euphoric,Relaxed	Earthy,Sweet,Citrus	NaN	NaN
2	13-Dawgs	hybrid	4.2	Tingly,Creative,Hungry,Relaxed,Uplifted	Apricot,Citrus,Grapefruit	Canadian LP Delta 9 BioTech	Canada
3	24K-Gold	hybrid	4.6	Happy,Relaxed,Euphoric,Uplifted,Talkative	Citrus,Earthy,Orange	NaN	Los Angeles
4	3-Bears-Og	indica	0.0	NaN	NaN	Mephisto Genetics	NaN

Original data frame

	flavor_1	flavor_2	flavor_3	flavor_4
0	Spicy/Herbal	Sage	Woody	None
1	Earthy	Sweet	Citrus	None
2	Apricot	Citrus	Grapefruit	None
3	Citrus	Earthy	Orange	None
4	NaN	NaN	NaN	NaN

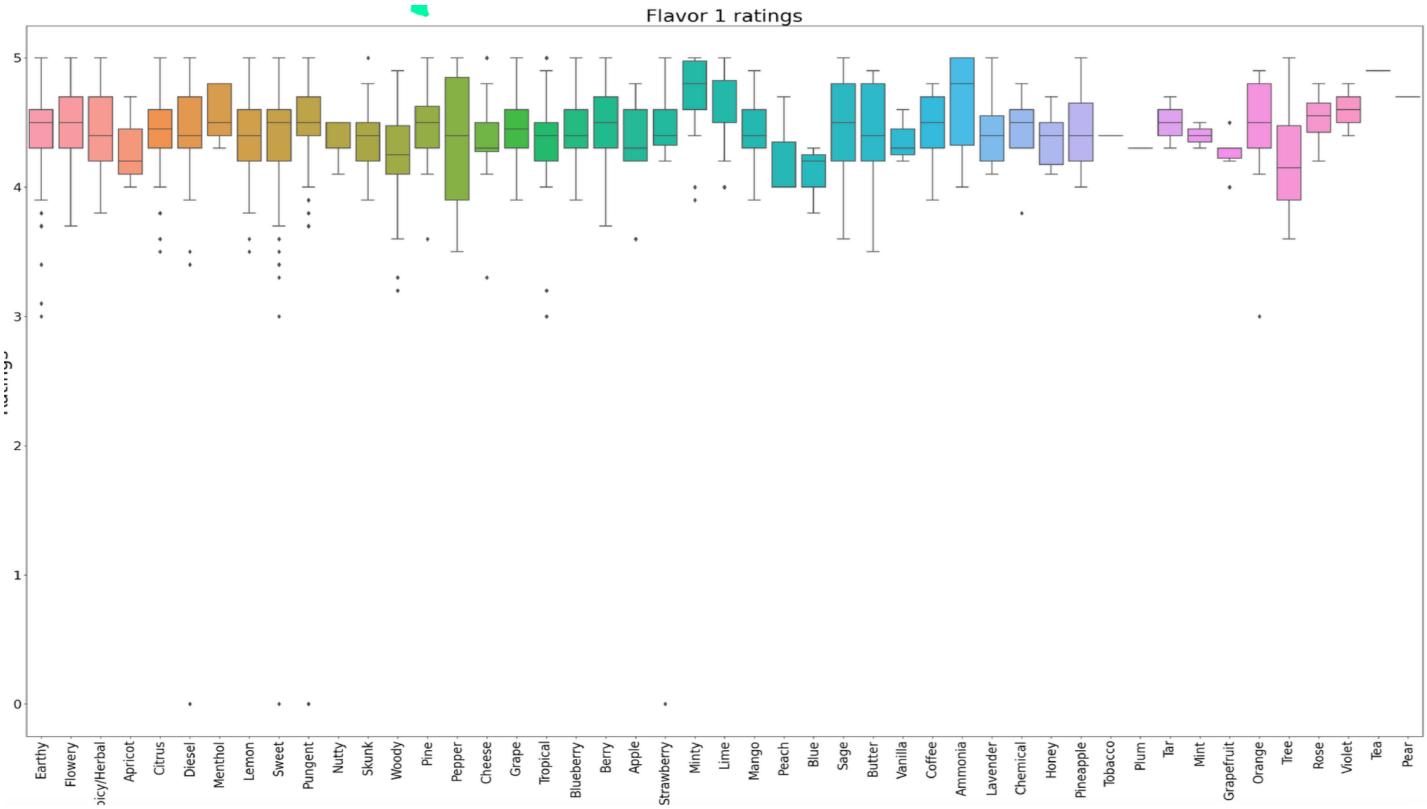
Flavors broken out to separate to columns from “Flavor” column in original data frame



	Strain	Type	Rating	Effects	Flavor	Breeders	LOCATION	flavor_1	flavor_2	flavor_3	flav
0	1024	sativa	4.4	Uplifted,Happy,Relaxed,Energetic,Creative	Spicy/Herbal,Sage,Woody	Medical Seeds Co	Spain	Spicy/Herbal	Sage	Woody	
1	100-Og	hybrid	4.0	Creative,Energetic,Tingly,Euphoric,Relaxed	Earthy,Sweet,Citrus	NaN	NaN	Earthy	Sweet	Citrus	
2	13-Dawgs	hybrid	4.2	Tingly,Creative,Hungry,Relaxed,Uplifted	Apricot,Citrus,Grapefruit	Canadian LP Delta 9 BioTech	Canada	Apricot	Citrus	Grapefruit	
3	24K-Gold	hybrid	4.6	Happy,Relaxed,Euphoric,Uplifted,Talkative	Citrus,Earthy,Orange	NaN	Los Angeles	Citrus	Earthy	Orange	
4	3-Bears-Og	indica	0.0	NaN	NaN	Mephisto Genetics	NaN	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	
2346	Zeus-Og	hybrid	4.7	Happy,Uplifted,Relaxed,Euphoric,Energetic	Earthy,Woody,Pine	A Greener Today's Dankcrar	NaN	Earthy	Woody	Pine	

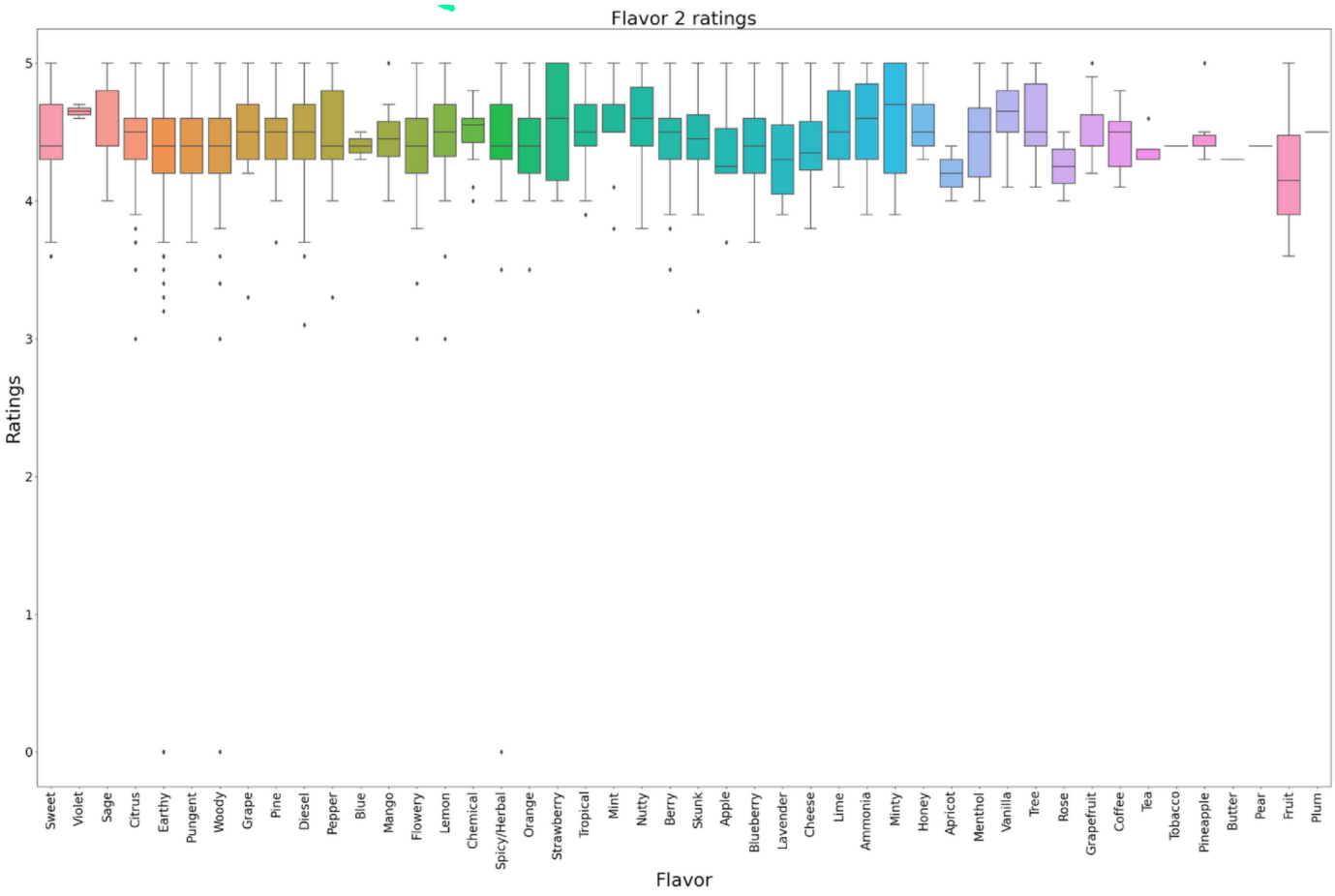
Combined broken out flavors into original data frame

Flavor profile one resulted in its ratings spanning across a range of 4.1 to 4.9. There were 47 flavors that made up this profile. The minimum rating went to a flavor called Blue and the highest rating was for Tea. The mean for all the ratings in flavor profile one was 4.43 and the mode was 4.40. The flavor profiles were large and the cleanest way to visually depict the data points per flavor would be to use a box plot chart. At a quick glance, the boxplot shows the minimum, first quartile, median, third quartile, and maximum values. Analyzing the graph for flavor profile one, the median lines line up horizontally across the board and there were a limited number of outliers. The y-axis shows the rating and the x-axis lists out each flavor.



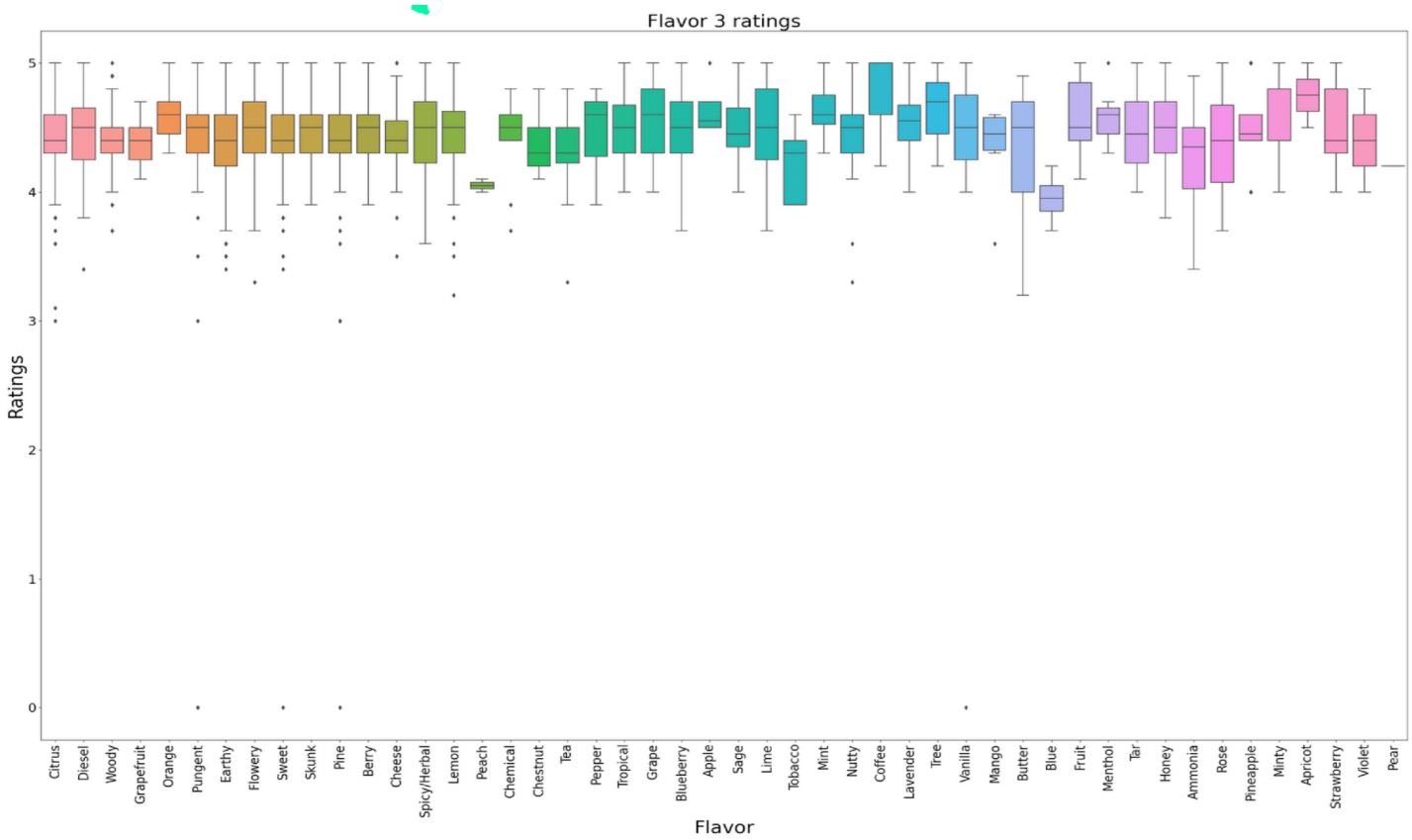
Box plot of flavor one profile

Flavor profile two had its ratings span across a range of ratings from 4.2 to 4.6. It encompassed 46 separate flavors. The minimum rating went to a flavor called Apricot and the highest rating was for Violet. The mean for all the ratings in flavor profile two was 4.45 and the mode was 4.50. The box plot of flavor profile two has a similar resemblance to flavor profile one.



Box plot of flavor two profile

Flavor profile three ratings spanned across a range of 3.90 to 4.70. It had a total of 48 flavors. The minimum rating went to a flavor called Blue and the highest rating was for Coffee. The mean for all the ratings in flavor profile three was 4.45 and the mode was 4.40. The box plot of flavor profile three has a similar resemblance to flavor profiles one and two.



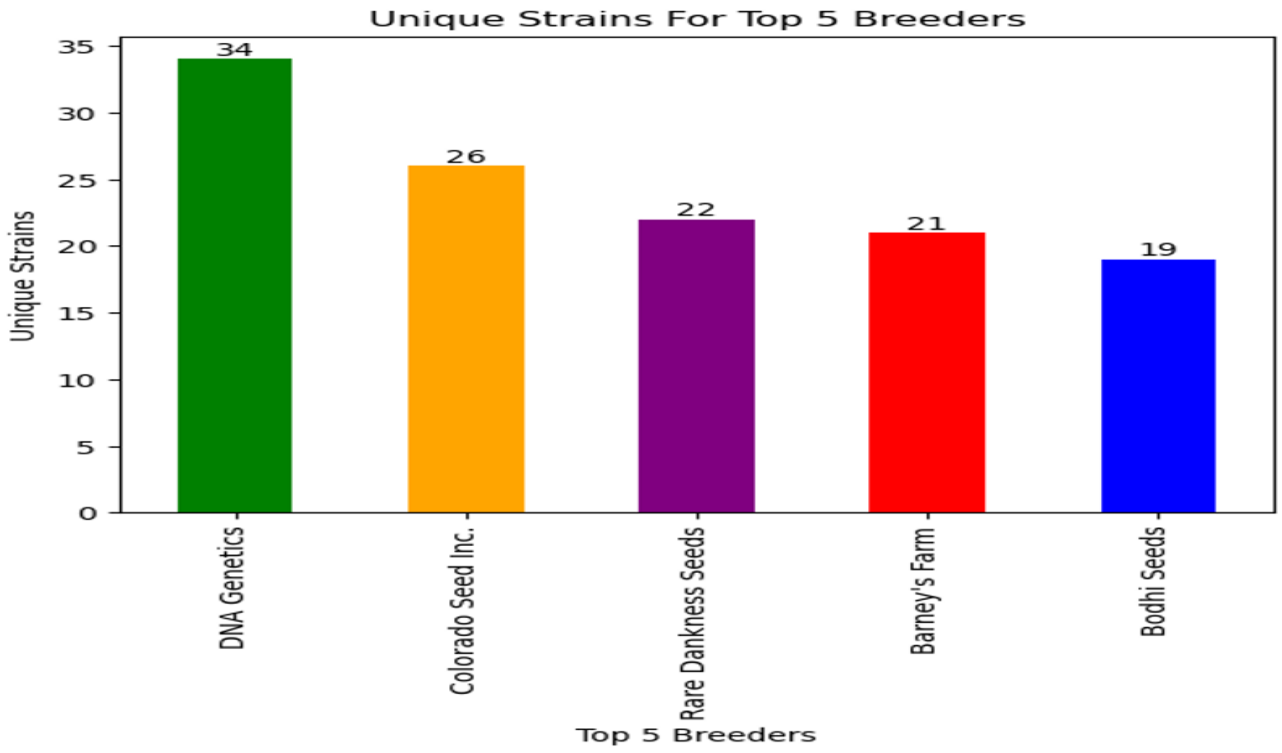
Box plot of flavor three profile

Comparing all three flavor profiles, the datapoint values are extremely similar with tight ranges, similar modes and means. This inclines Team 2 to trust comparing values across all the flavors because they are comparable datasets. Each flavor profile was roughly equal, giving a high confidence level that the data samples are similar enough to compare for statistical analysis.

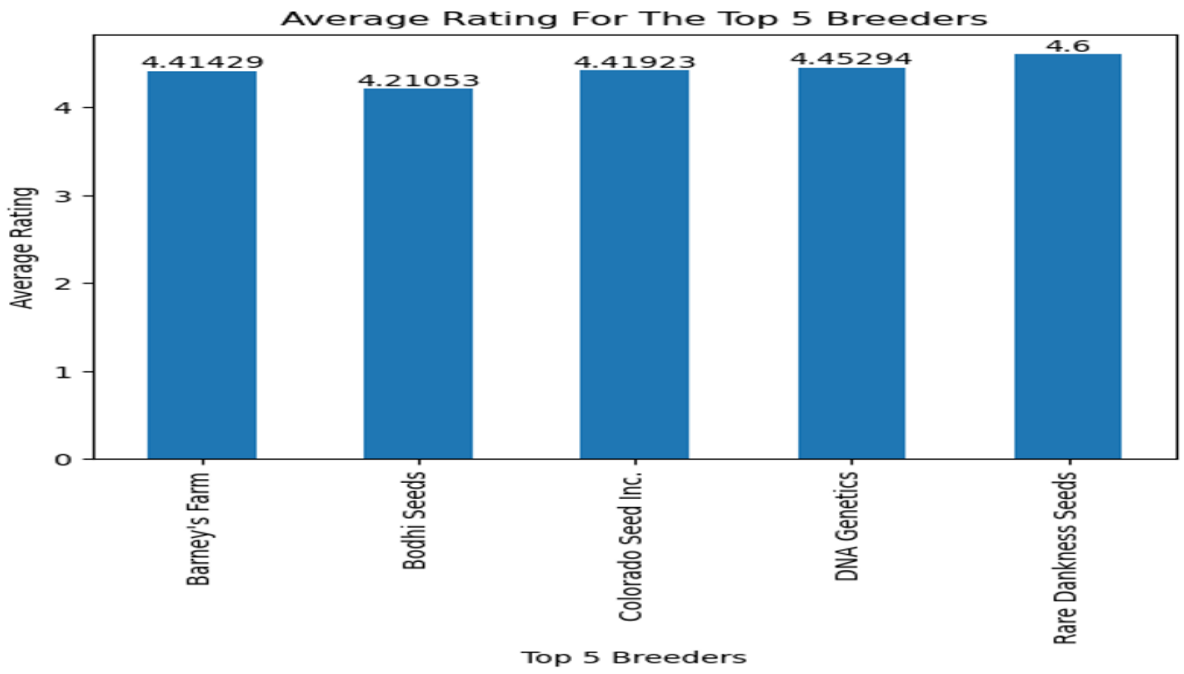
In conclusion, hypothesis number 2 was incorrect with assuming sweet as having the highest rating. The team was surprised to find that Tea had the highest rating with Coffee closely behind it. One would not naturally associate cannabis flavors being popular for mimicking beverage flavors, but coffee and tea are a national favorite so maybe consumers are accustomed to those flavors. The team would want to explore if this result would be seen globally as well since coffee and tea can be found worldwide. Future research of this dataset would include grouping the flavors into larger groups for analysis. Grouping flavors by the five different types of taste on the human tongue receptors such as salt, sweet, sour, bitter, and umami would allow for more general analysis on what flavors are preferred by the population, nationwide and globally.

Hypothesis 3: Discussion and Analysis

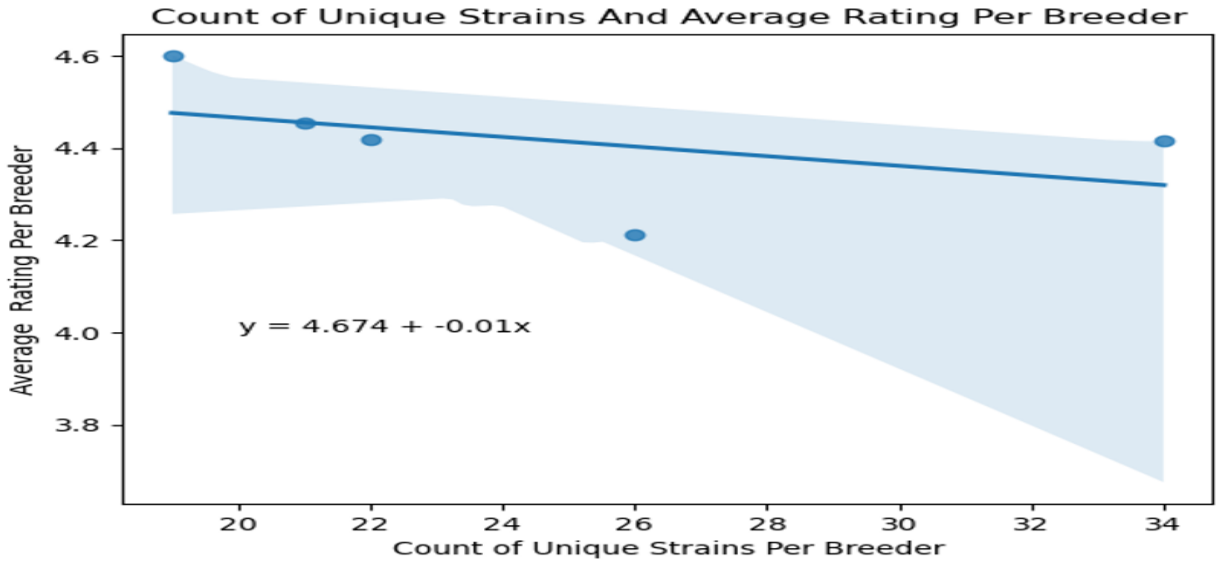
After reviewing the data, we were curious to know who the top breeders are and what types of cannabis they produce. We determined the Top 5 Breeders by counting each of their unique strains.



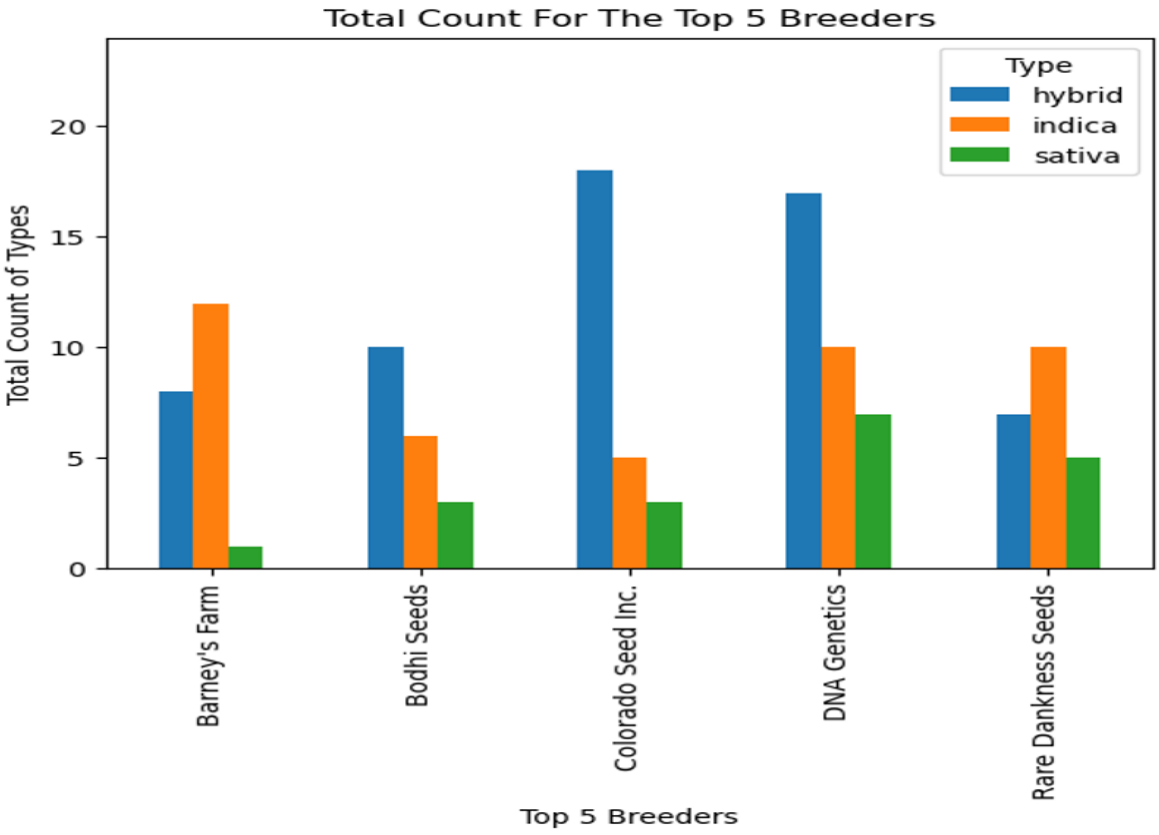
This led to hypothesis #3, Do breeders with the highest count of unique strains have the highest average ratings? We found 40% of the Top 5 Breeders had the same ranking based on the quantity and average rating of their unique strains.



Another finding is 60% of the Top Breeders had no significant correlation in these two categories. After analyzing the present data, we found there is no significant correlation between the count of unique strains and average rating per breeder. For example, Rare Dankness Seeds came in 1st place for the highest average rating and 3rd place based on their total number of unique strains.



An additional and surprising takeaway discovered in this dataset: 60% of the Top 5 Breeders' most popular cannabis type is hybrid, while 100% of the Top 5 Breeders least popular type of cannabis is sativa. In earlier research we uncovered the main effect for sativa is "happy!"



Hypothesis 4: Discussion and Analysis

The average rating by Type (hybrid, Indica, and sativa) is almost equal in all Types. See the code below to see the mean, min and max of all Type. The average of the Types is between 4.29 and 4.30; and 4.91 and 4.95.

```
#Find Average Rating for each Type
#What is the average(mean, min, max) Rating reported by Type?
OD2.groupby("Type").Rating.mean()
x = np.linspace(0, 5, 10)
y = np.sin(x)
plt.plot(x, y, '-ok', color='green', marker='*');
plt.title('AVERAGE RATING BY TYPE', color = 'BLUE', fontsize = 16)
plt.xlabel('EFFECT_1', color = 'BLUE', fontsize = 16)
plt.ylabel('RATING', color = 'BLUE', fontsize = 16)
print(OD2.groupby("Type").describe())
```

	Rating								num_effects
	count	mean	std	min	25%	50%	75%	max	count
Type									
hybrid	1212.0	4.291667	0.885812	0.0	4.2	4.4	4.7	5.0	1163.0
indica	699.0	4.347783	0.750954	0.0	4.2	4.4	4.7	5.0	680.0
sativa	440.0	4.303864	0.824847	0.0	4.2	4.4	4.6	5.0	421.0

	mean	std	min	25%	50%	75%	max
Type							
hybrid	4.914015	0.470655	1.0	5.0	5.0	5.0	5.0
indica	4.925000	0.410946	1.0	5.0	5.0	5.0	5.0
sativa	4.952494	0.327508	1.0	5.0	5.0	5.0	5.0

The average rating by effect_1 (total 14) is almost equal in all Types. See the code below to see the mean, min and max of all effects in effect. The Dry effect has the lowest mean at 4.00. They only had a rating count of one (1). The Creative effect has the highest mean at 4.47. They had a rating count of 80.

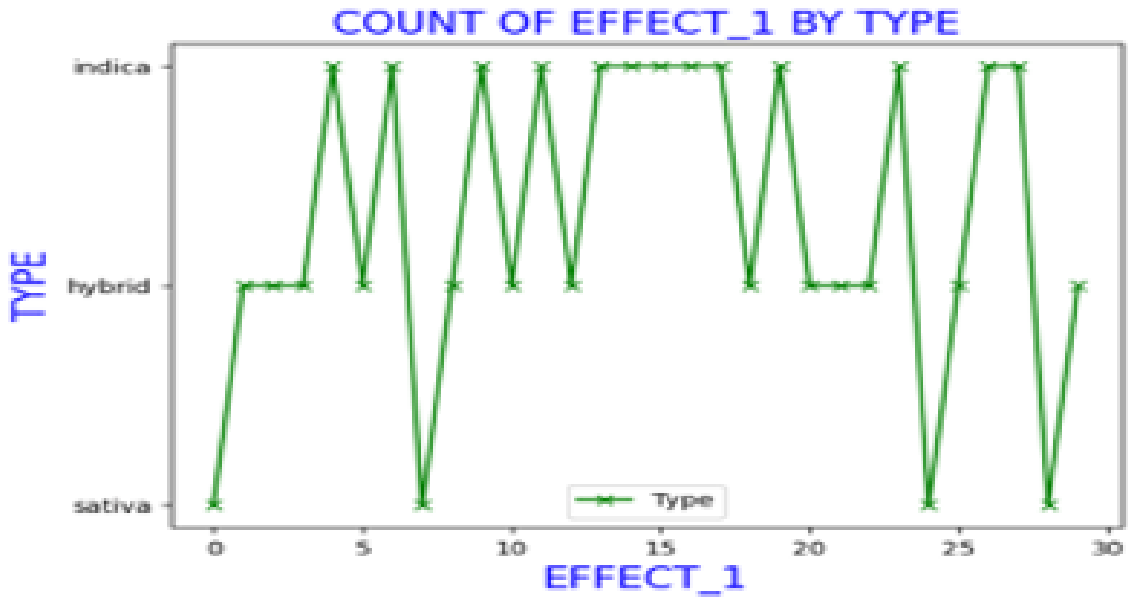
```
#Find average rating for each effect_1
#What is the average(mean, min, max) Rating reported by Effect?
OD2.groupby("effect_1").Rating.mean()
print(OD2.groupby("effect_1").describe())
```

```
           Rating                                                          \
            count      mean       std  min   25%   50%   75%  max
effect_1
Aroused      38.0  4.450000  0.904269  0.0  4.425  4.70  5.000  5.0
Creative     80.0  4.473750  0.611854  0.0  4.275  4.50  4.800  5.0
Dry           1.0  4.000000       NaN  4.0  4.000  4.00  4.000  4.0
Energetic    74.0  4.332432  0.578877  0.0  4.200  4.40  4.600  5.0
Euphoric    249.0  4.425703  0.518510  0.0  4.200  4.40  4.700  5.0
Focused      53.0  4.366038  0.377718  3.4  4.200  4.40  4.700  5.0
Giggly       12.0  4.233333  0.986577  2.0  3.900  4.60  5.000  5.0
Happy       476.0  4.412395  0.297346  3.0  4.300  4.40  4.600  5.0
Hungry       36.0  4.125000  1.100487  0.0  4.000  4.30  4.725  5.0
Relaxed     825.0  4.448364  0.272618  3.0  4.300  4.50  4.600  5.0
Sleepy       89.0  4.291011  0.916223  0.0  4.100  4.40  4.800  5.0
Talkative    55.0  4.376364  0.822364  0.0  4.200  4.50  5.000  5.0
Tingly       32.0  4.331250  1.226176  0.0  4.275  4.75  5.000  5.0
Uplifted    244.0  4.468852  0.458227  0.0  4.300  4.50  4.700  5.0
```

Without speaking to the consumers and reviewing just the dataset, we can clearly see the population perceives effects differently when consuming the same type of cannabis.

```
#Visualization Count for each Effect for effect_1 by Type
OD2.groupby(["Type", "effect_1"]).size()
plt.plot(OD2['Type'][:30], marker = 'x', color = 'green', label = 'Type')
plt.title('COUNT OF EFFECT_1 BY TYPE', color = 'BLUE', fontsize = 16)
plt.xlabel('EFFECT_1', color = 'BLUE', fontsize = 16)
plt.ylabel('TYPE', color = 'BLUE', fontsize = 16)
plt.legend()
```

```
<matplotlib.legend.Legend at 0x297e14730>
```



CONCLUSIONS

In conclusion, the analysis of our dataset provided valubale insights into the types, effects, flavors, breeders, and consumer ratings of cannabis; however, it is equally important to highlight the limitations and boundaries we encountered in order to refine the scope of our analysis and create a springboard where future cannabis research can begin. We focused our statistical analysis on the first three flavors listed in the raw data, prioritizing the order in which the effects and flavors were presented. For instance, in the dataset, there were 5 or more effects listed in one single cell, therefore, we used the first of the five effects listed to evaluate and draw conclusions for our analysis, and so on.

Despite the abundance of potential information available to glean insights from the data's "description" column, our technical capabilities in the area of advance programming and language processing was limited. Moreover, the dataset was exclusive to legal breeders, meaning that all non-regulated breeders and consumer data were not included. Given additional time and programming resources, we would enhance the data extraction process to include award-winning cannabis varieties and their associated growers listed in the "description" column of the dataset.

While our data cleaning process provided us with a robust total amount of 2351 reviews per strain and sample sizes, the dataset did not include the age ranges of the consumers nor their state of mind while reviewing the product. Consequently, we do not have enough conclusive evidence to determine whether a cannabis type designed to elicit a "happy" effect caused the validity of the ratings data to be skewed due to implicit bias. In future research, a questionnaire or consumer interview would be potential option implemented to address these nuances and provide a more comprehensive understanding of the specific facets at play given these circumstances as they relate to cannibas use.

Now that we have addressed the limitational boundaries inherent from this cannabis dataset, we can use the aforementioned constraints to serve as a guidepost in understanding the meaningful implications as it relates to the context of cannabis on a larger, in-depth scale. The implications of our cannabis analysis spans across various sectors including government, health and medical, consumers, and breeders. For instance, the advancements and changes that can occur in government includes continued refinement of regulatory guidelines when defining medical and recreational use of cannabis for consumer and breeders. Next, healthcare professionals can recommend specific cannabis strains and types to treat mental health and mood disorders, particularly as more research studies are conducted under legalized varieties of cannabis. In addition, consumers can be confident in making informed choices based on ratings, experiences, and trusted breeders. Finally, breeders can educate consumers about their brand, ensure consistency, as well as predictability in the quality of their product. As an

advantage, the insights of these implications can be used to transform future data driven decisions as the legal cannabis landscape continues to evolve in a safe and positive direction.

To summarize, the main objective of our project was to conduct a comprehensive analysis using a cannabis dataset, incorporating a variety of factors such as type, flavors, moods, effects, breeders, strains, and user ratings to highlight comparative insights and relational patterns amongst the production and consumption of cannabis. The popularity of cannabis is on an steep upward incline, and there is a noticeable shift in attitudes and beliefs surrounding the medicinal and recreational use of legal cannabis, particularly in the United States. Recent research efforts indicate that professionals in various industries are working to successfully bridge our current understanding of the potential positive and negative impacts cannabis can have on our society, government, healthcare, and businesses. Even though the resounding query of “Do You Love Me, Mary Jane?” still stands as a rhetorical question for the past 45-years straight, those who have an invested, yet consumable interest in the current landscape of legal cannabis will no doubt be in a state of relaxation, happiness, or euphoria as they perpetually await for a reply from Ms. Mary Jane.