Sahir Doshi, Sarina Zaparde, Joonyoung (Jase) Jeon, and Qinglin (Jason) Yang

BA222

Leder-Luis

13 November 2022

Project 2: Analysis of Cannabis Strain Effects

The dataset we are using examines the physical and psychological effects of various cannabis strains on users. Since cannabis is legal in Massachusetts, we felt that it would be informative and helpful to explore the potential health benefits it has on users. Our data came from Kaggle[1] where the creator scraped the original information from Leafly—a website dedicated to providing free and thorough cannabis education. The data is publicly sourced from user feedback through the Leafly website where the feedback percentages are based on buyers' reviews. Questions we hope to answer through our analysis include what medical issues the cannabis strains help solve the most/least, the impact of THC level (the psychoactive component) on users, and differences caused by varying primary chemical structures.

The dataset tracks the effect of cannabis strains on positive emotions, negative emotions, and pain relief for diseases/health conditions based on THC level for each cannabis strain. The strains are further categorized into types (Indica, Sativa, or Hybrid) and most common terpene—a naturally occurring chemical compound—within each strain. The percentage value for each emotion and disease indicates how many users of the specific strain reported that cannabis consumption helped them with the emotion/condition they are having (ex. 50% for pain relief means that half of the users for this cannabis strain reported it helped them with pain relief).

We cleaned up the data by removing all *NaN* observations in *thc_level*, *most_common_terpene*, *happy*, and *stress* variables. The format for *thc_level* was converted from percentages to decimals for easier regression analysis. Strains with *Humulene*, *Linalool*, and *Ocimene* terpenes were discarded as the dataset (lack of data). The *img_url* and *description* were

[1] https://www.kaggle.com/datasets/gthrosa/leafly-cannabis-strains-metadata. Accessed Nov. 12 2022.

also unused as they do not contribute to our analysis. Finally, our data is 2 years old, so it is possible that percentages for the cannabis strains have changed since then.

	thc_level	happy	stress	lack_of_appetite	nausea	headaches	cramps	inflammation	muscle_spasms	eye_pressure
count	2420.000000	2420.000000	2420.000000	2420.000000	2420.000000	2420.000000	2420.000000	2420.000000	2420.000000	2420.000000
mean	0.185293	0.394678	0.231376	0.030298	0.014260	0.026562	0.006616	0.019405	0.014202	0.006583
std	0.035045	0.280550	0.204575	0.101229	0.072442	0.097662	0.047491	0.088940	0.073106	0.052677
min	0.060000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.170000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.180000	0.480000	0.260000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.200000	0.590000	0.370000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	0.340000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Figure 1. Summary statistics of the dataset

Because each strain has a different number of users which directly affects the sample size of reported effects on them, it is possible that the data for them are unreliable and not representative of the true population. Furthermore, to ensure that our data is reliable, we only examined variables that had >30 nonzero values so that our findings are valid, leading to 2420 observations and 62 variables. As seen in Figure 1 above, which displays the summary statistics of our dataset (note only the first couple variables are displayed for the interest of space), all the cannabis strain's best effects were helping users feel more happy and less stressed, as those 2 emotions had the highest average percentage. As for conditions like muscle spasms and eye pressure that had extremely low percentages of 1.42% and 0.66% respectively, it is clear that cannabis was less effective in helping users resolve those issues.

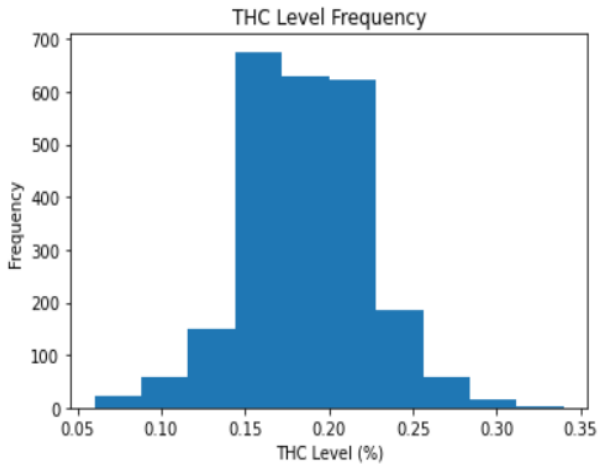

Figure 2. Histogram showing THC level frequency

As seen in Figure 2, the wide majority of strains recorded fell within a 15% - 20% THC level overall. Because we wanted to assess the effects of THC concentration on how it relieves specific medical conditions or physically affects users, the distribution of values that fall outside the center range provided us with enough data points to predict different outcome variables using different regressions.

	thc_level			happy			stress		
	mean	std	count	mean	std	count	mean	std	count
type									
Hybrid	0.186691	0.034959	1384	0.417962	0.287254	1384	0.228634	0.209130	1384
Indica	0.181595	0.032673	489	0.441963	0.185526	489	0.318609	0.173902	489
Sativa	0.180764	0.032888	288	0.533125	0.202746	288	0.294097	0.146399	288

Figure 3. GroupBy for cannabis type

Based on the GroupBy in Figure 3, for the THC levels and the effects of each strain on happiness by type, we can see that Hybrid is the most popular type at a count of 1384, followed by Indica at 489 and Sativa at 288. Hybrid has the highest average THC level but also the lowest percentage for the *happy* variable, meaning that despite having a higher THC level, less users reported Hybrid strains help them feel more happy. On the other hand, Sativa has the lowest average THC level but had a higher percentage of users report it helping them feel more happy. Although THC level may have affected these results, it is also important to note that Hybrid strains in general help with relaxation, whereas Sativa strains help more with anti-anxiety benefits.

	thc_level			happy			stress		
	mean	std	count	mean	std	count	mean	std	count
most_common_terpene									
Caryophyllene	0.188296	0.031916	452	0.381128	0.283616	452	0.204358	0.203570	452
Limonene	0.197601	0.034569	371	0.360485	0.291107	371	0.209326	0.217278	371
Myrcene	0.180268	0.033937	1195	0.414887	0.269126	1195	0.258603	0.194311	1195
Pinene	0.173800	0.040745	100	0.421700	0.270155	100	0.247700	0.221404	100
Terpinolene	0.189371	0.037407	302	0.368046	0.303718	302	0.185762	0.208574	302

Figure 4. GroupBy for most common terpene

In Figure 4, shown above for the GroupBy for the most common terpenes, we removed 3 other terpenes from the dataset due to them having less than 30 observations as well as any cannabis strain left uncategorized for the terpene. As shown in the GroupBy, Myrcene was by far the most common terpene and also had one of the highest average percentage scores for both the *happy* and *stress* variable, indicating that users felt that it was the most helpful terpene to benefit them. Pinene also had high average percentage scores for both variables, but interestingly enough was the least used terpene and only appeared in 100 strains. This could be due to the fact that since it has greater effects, it is more expensive to buy and therefore less users are willing to pay for it.

Dep. Variable:	lack_of_appetite	**R-squared:**	0.001
Model:	OLS	**Adj. R-squared:**	0.000
Method:	Least Squares	**F-statistic:**	1.160
Date:	Sun, 13 Nov 2022	**Prob (F-statistic):**	0.324
Time:	21:51:52	**Log-Likelihood:**	2111.1
No. Observations:	2420	**AIC:**	-4214.
Df Residuals:	2416	**BIC:**	-4191.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0140	0.011	1.239	0.215	-0.008	0.036
thc_level	0.0786	0.059	1.335	0.182	-0.037	0.194
Indica	0.0072	0.005	1.384	0.166	-0.003	0.017
Sativa	0.0022	0.006	0.333	0.739	-0.011	0.015

Figure 5. Regression for percentages by type

Dep. Variable:	lack_of_appetite	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	-0.000
Method:	Least Squares	**F-statistic:**	0.7601
Date:	Sun, 13 Nov 2022	**Prob (F-statistic):**	0.579
Time:	21:51:58	**Log-Likelihood:**	2111.3
No. Observations:	2420	**AIC:**	-4211.
Df Residuals:	2414	**BIC:**	-4176.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0213	0.012	1.742	0.082	-0.003	0.045
thc_level	0.0593	0.060	0.991	0.322	-0.058	0.177
Limonene	0.0035	0.007	0.491	0.624	-0.010	0.017
Myrcene	-0.0031	0.006	-0.559	0.576	-0.014	0.008
Pinene	-0.0105	0.011	-0.937	0.349	-0.033	0.011
Terpinolene	-0.0044	0.008	-0.589	0.556	-0.019	0.010

Figure 6. Regression for percentages by terpene

As seen in Figure 5—for the regression analysis we did to predict how lack of appetite would be affected by changes in the variables of THC level and type—the dummy variables Indica and Sativa have coefficients of .0072 and .0022 respectively, indicating how much the percentage for lack of appetite would increase when the cannabis strain is one of these 2 types. The THC level coefficient of .0786 indicates how much the percentage for lack of appetite would increase when the THC level increases by 1.

Figure 6 shows the regression analysis predicting how the percentage for lack of appetite would change based on THC level and most common terpene changes. The terpene coefficients indicate how much the percentage would increase or decrease when a cannabis strain is of a specific terpene, and the THC level coefficient of .0593 indicates how much the percentage would increase when the THC level increases by 1.

Both Figure 5 and 6 have very low R-squared values of .001 and .002 respectively, indicating that the percentage of users that report using cannabis strains helps with lack of appetite is unaffected by the THC level, type of cannabis, and most common terpene, so there is no correlation between these variables and their effects on helping treat lack of appetite more.

With the given information, we cannot establish a direct causal relationship between THC level and any of the other variables (such as dizziness) as there is omitted variable bias. This includes information on how tolerant each respondent is to cannabis based on prior use, medications taken, and other health conditions which could exacerbate the effects of THC and dizziness, for example. Furthermore, we cannot establish a causal relationship either with “type” or “most common terpene” either, because there is selection bias for users who may use cannabis more heavily or have pre-existing health conditions.

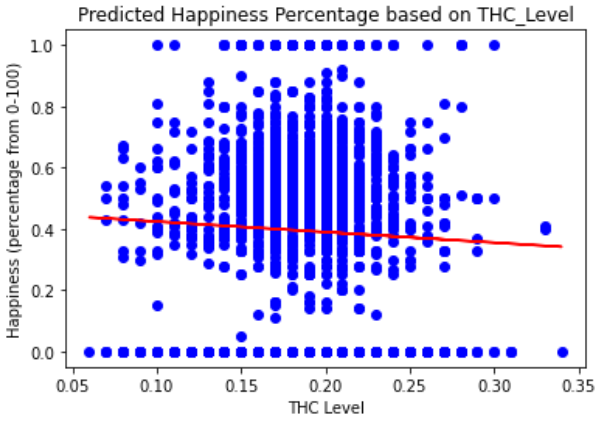

Figure 7. Multivariate Graph of happiness by THC level

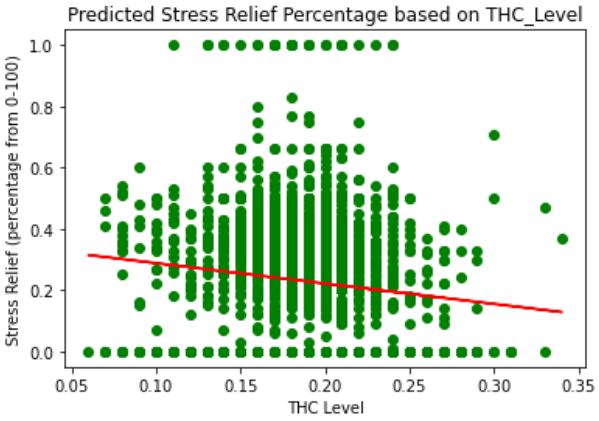

Figure 8. Multivariate Graph of stress relief by THC level

One may hypothesize that higher THC levels will lead to a higher degree of happiness. However, looking at Figure 7—a scatter plot between different strains and percentage of users indicating the strain was effective in increasing their happiness level—we discovered that this is not necessarily the case, as there is a negative correlation of -0.3437 between the two variables. Likewise, the percentage of users indicating the strain helped with reducing stress levels decreased in a coefficient of -0.6652, indicating a quicker drop in effectiveness than in helping with happiness. Regardless of scientific truth, a subjective response of users indicates that there is a steady decrease in the level of happiness the strain brings for strains of THC level higher than .20.

Contributions

Sahir Doshi focused primarily on conducting analysis while coding, acting as the primary programmer. Sahir created the following sections in the Jupyter Notebook: Data Cleansing, Regressions, Multivariate Graphing (and also organized and formatted the code including descriptions and headers in Markdown blocks). Jason Yang assisted in programming by completing the following section: Basic Statistical Analysis, and Jase Jeon also made a few minor edits to the code. Sarina Zaparde focused primarily on the report, writing most of the sections, while Jason and Jeon helped Sarina as well, specifically writing on the interpretation of certain graphs and/or tables. Finally, all 4 group members participated in researching datasets and subsequently agreeing on a viable one, several discussions debating the direction of our analysis (e.g. which variable to predict, what data to remove), and editing the report.