Predicting THC from Bud Apogee

February 2020

CSE 163 Project Report

Research Questions:

1) **How accurately can we predict the THC level in a legally grown strain of cannabis?**
   We'll be analyzing a wide range of variables/factors in cannabis production to try and determine the level of THC content within the strain. Some of these variables include the type of strain, chemotaxonomy, cultivation methodology, etc.
   Answer: We can predict THC levels somewhat accurately. Our values are in general proximity of the actual value with not too much error for the most part (more in the Results section).
2) **Are some variables more important than others in determining THC content (or, do other variables possess a stronger correlation with THC levels)?**
   It's important to, early in our research, rule out any variables that may have no effect on THC content. The way we'll handle this will be explained in the Methodology section of this proposal. Also, not only will we remove extraneous variables, but we'll also work to come up with a sort-of "ranking" of each factor and decide on which has a stronger correlation with THC content (CBD level vs. cultivation methodology, for example).
   Answer: Yes, some variables play a greater role in determining THC content.
3) **After predicting THC levels for specific strains, how does understanding their respective attributes allow us to predict their popularity in the marijuana community?**
   While the focus of our project and algorithm is on the previous research questions, we still believe it's necessary to find out if there exists a correlation between the potency of a particular marijuana strain and its reputation within the marijuana community. We'll discuss this further in the Motivation section, but answering this question is of great importance within the discussion of potential marijuana legalization.
   Answer: Initially, we assumed the statistic we are provided in the dataset for Leafly reviews was going to be a score rather than a ranking. For example, we thought certain strains would have values like "4.3/5" and "3/5" so comparison between strains is easier. However, we're provided with a ranking where the highest-rated has a value of "1" and the worst would be in the hundreds (if not thousands). This fact makes it harder to answer this research question because it's hard to find a correlation between THC content and the Leafly review rankings with a simple regression model. More on this in the Results section.

Motivation & Background:

According to the Marijuana Investigations for Neuroscientific Discovery program at Harvard, there's been a considerable increase in the potency of marijuana (THC) from 1995 to now. More specifically, THC levels have risen approximately 300% since 1995. They also state, "The negative effects of cannabis have primarily been isolated and localized to THC…it stands to reason that higher levels of THC may in fact confer a greater risk for negative outcome." As marijuana users build their tolerance over-time, they begin to chase "stronger highs" (Appendix D) and this raises concerns about health risks. High concentrations can cause panic attacks and cause users to feel psychotic effects and paranoia, and can produce massive vasoconstriction, which leads to decreased blood flow through one's vessels.

So, what does this all mean for our particular project? First and foremost, many recreational marijuana users might not be aware of the exact THC content of the cannabis they smoke. Producers are required to have a THC percentage label on their products, sure, but this is usually small and can go about as unnoticed as obscure ingredients on food labels. Also, since marijuana is illegal under federal law, many people purchase this drug through illicit means (meaning they have no idea about the exact potency). If we can create an easy-to-use algorithm that can spit out good approximations of THC level for just about any strain, users are better informed as to what they are smoked and this promotes safer use of the drug. For example, when an individual purchases flower "off the street", they generally only know the amount they are purchasing and the type of strain as these are the two biggest determinants in the price of marijuana. They would be able to use our algorithm in order to find out how potent that particular strain actually is so they aren't in the dark regarding what they're consuming.

Another reason why this is worth computing is for the use of marijuana production facilities and farms. Producers are more likely to make a better product for their consumers if they are better aware of what variables matter the most in creating less or more potent marijuana. Also, this would increase efficiency in production with less testing and will subsequently lower prices on their products.

Lastly, it's important to understand how a marijuana strain's popularity might reflect its potency so that researchers, scientists, and lawmakers alike can use this information to better assess the risks of the ever-increasing popularity of marijuana as a whole. If we find that strains with lower THC levels tend to be more popular, a stronger case for the federal legalization of marijuana can be made. However, if we find the opposite to be true, this information could be used be possibly implement laws that limit THC content in recreational marijuana.

Dataset:

Link: https://www.dolthub.com/repositories/Liquidata/cannabis-testing-wa/data/master/tests

This dataset is composed of over 200,000 laboratory measurements of cannabis products for legal sale in Washington state. It includes 21 columns that identify different aspects of each strain, ranging from their chemotaxonomy to their user reviews on a popular cannabis website. Not all of these columns will be of-interest and we'll discuss this more within our methodology.

Methodology:

To begin our data analysis, we'll work to clean up our dataset (using pandas) and remove any extraneous columns from the CSV (namely 'org_active', 'date_test', 'strain_leafly_page_rank' among others). Also, we'll drop any rows with missing values. With a consolidated dataset consisting of only the columns we are interested in, we'll be ready to begin our testing to see if any of these columns may not have any effect on THC content. In other words, the specification process for our regression begins. The dependent variable is THC content, with everything else being independent (multiple regression model). For this, we'll perform hypothesis testing with different null hypotheses for each of our independent variables where we see if their true beta value is equal to zero (traditional t-test). We'll then compare the t-scores to critical values for our respective sample size. A combination of Python's SciPy and Numpy libraries will be necessary to perform these tests (possibly Statsmodels as well) as they have built-in functions for regression analysis.

However, it's important to bear in mind that there are two questions to answer in regards to model specification. The aforementioned procedure deals with whether the independent variable is statistically-significant or not, but an equally important question to ask is whether the variable is essential to the regression on the basis of theory and that's why we perform the data clean-up mentioned at the start of this section. No matter how high the test-statistic or a correlation coefficient is, we'll remove any variables that simply aren't

theoretically-sound (for example, whether the producer of the strain is still operating or not is an irrelevant variable as it contributes nothing theoretically).

Once we've completed the test for extraneous variables, we'll begin on building our model for Research Questions #1 and #2 using Sciki-Learn (the golden standard for ML in Python) and its various modules. We'll split our dataset into two different data-frames, one just containing the dependent variable and the other with all the independent variables. With this, we're able to fit our model and use it to make predictions. In order to make strong predictions, however, we won't use the entire dataset but rather we'll split our data into training and test data. The fact that our dataset contains over 200,000 data points makes the process of testing our model and its accuracy significantly easier. We'll need to be careful to not let our test set be too big, as we'll start lacking data to train on. We'll pick between 15-25% of the total data to make up our test set (the rest being the train set).

After coding the machine learning portion, we'll want to visualize the train and test sets and see if their plots look relatively the same (using the seaborn and matplotlib libraries), or trend in the same direction. We'll first build a scatter plot for each and use sci-kit's LinearRegression() function to draw a line over each plot (in a different color than the points) in order to evaluate them effectively. At this point, we should be able to make predictions for THC content based on different values for each of the independent variables, and determine which variables hold a stronger correlation with the dependent variable.

For the last research question, regarding strain popularity, we'll use a simple linear model with one dependent variable (Leafly rating from 0.0-5.0) and one independent variable (THC content) in order to determine if there exists a correlation. The model we create for this particular research question will be separate as it isn't theoretically-sound for a strain's rating on a review site to have an effect on its THC content (we are testing if it is the other way around).

Results:

Note: Data visualizations that don't involve machine learning are all viewable in the Appendix at the end of this paper. Whenever a graph/chart is mentioned, we reference which part of the Appendix you'll find it.

The main question we wanted to answer with this project is 'how accurately can we predict the THC content in a strain of cannabis?' However, before answering that question, we decided we first wanted to determine which variables actually played a role in the amount of THC content in a strain. After ridding the data of non-theoretically-sound columns (insignificant variables), we used the Statsmodels library in order to run an ordinary-least-squares basic regression with each independent variable left in the data.

Do some variables play a larger role in determining the THC content in a strain of cannabis?

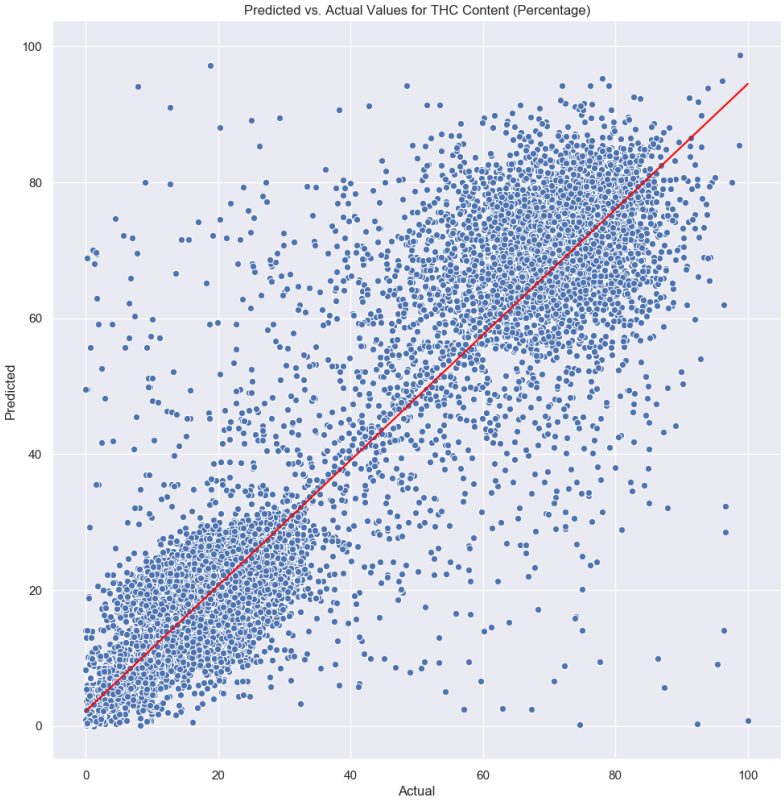
From the OLS technique, we were able to determine that the most significant variable was the cultivation methodology (or, 'inventory_type' as its called in the dataset). This pertains to what form the cannabis is in (hash, ultra-refined for dab pens, flower, etc.). Approximately 80% of the variation in THC content is explained by the cultivation type. This makes sense theoretically because if you go to a dispensary, you'll notice that the cannabis used in dab pens, for example, are known to have extremely high THC content (wax form) whereas edible forms of cannabis will have much less. The form in which the cannabis is sold has a direct relationship with THC content. A bar graph highlighting the different forms and their average THC content shows this well (Appendix A).

We found that chemotaxonomy (chemical make-up of the plant) explained about 60% of the variation in THC content, meaning we can predict THC from it only moderately well. This is interesting because the graph that shows THC content for each chemotype (Appendix C) makes it apparent that there is a significant difference in THC levels between each chemotype. However, the correlation coefficient is telling us that you can't strictly use the plant's chemotype to determine THC content (it would need to be paired in a multi-regression model).

Both CBD level and Strain Type (sativa, indica, etc.) explain pretty much none of the variation in THC content, and this makes a lot of sense. Regarding strain type, its different cannabinoid composition that sets indica apart from sativa while THC is a shared cannabinoid that exists in all weed strains. There's no reason for the potency to differ between different strain types. When marijuana is produced for purchase, the producers market it as either CBD or THC. So, it's expected that the CBD products/strains will have less THC since the buyers for that product will not want any. The producers ensure that there is less CBD with products marketed as THC and vice versa. However, CBD content is low on average and sometimes very slight amounts (.01%) can appear in strains and this makes it statistically difficult to find a correlation.

How accurately can we predict the level of THC content in a strain of cannabis?

Moderately accurately, for lack of a better term. Our machine learning model produces a mean squared error (MSE) of ~51. This means that our error, on average, was roughly 7 with THC content being valued at a range of 0-100. This isn't very good, but it's not bad either. Sometimes we might have a predicted value of 23% THC content with the actual value being 18%, but other times a predicted value could be 15.8% and the actual value is 15.9%. It's safe to say that our model predicts near-perfectly about half the time. The model in question uses every theoretically-sound independent variable, and it's clear that 'inventory_type' is doing the heavy-lifting. It's worth noting that when building the model with just 'inventory_type' as the independent variable, it yields worse results. Every independent variable in the restricted dataset was necessary to produce the best possible model. Our graph for the model's accuracy, fit with a line, shows a decent relationship between our actual and predicted values (shown on the right).

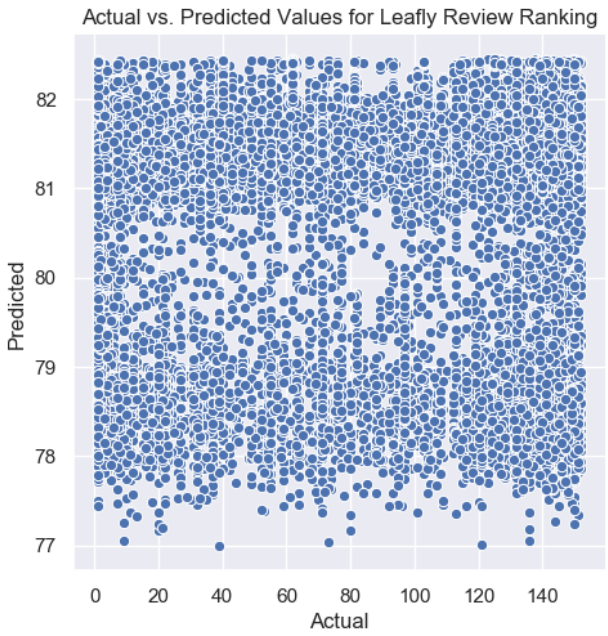


The result from this model was honestly expected from the both of us. It's quite the tall and daunting task to be able to perfectly predict cannabis potency using only the variables we had in the dataset because there is so much that goes into the production of cannabis on the chemical-side of things. It's hard to include those factors in a spreadsheet format, so our model was limited from the beginning.

An interesting aspect of the graph on the right shows what we know about cultivation type's effect on THC content. There is a large cluster of data points on the bottom left, and an even larger one on the top right. This is because most of the cannabis sold at dispensaries, and therefore most of the cannabis in this dataset, is either in flower or wax form (for joints/pipes/blunts and dab pens respectively, the most popular mediums people use to smoke marijuana).

Does THC content of a particular strain have an effect on that strain's popularity?

The short answer? No. As we discuss in the Research Questions portion of this paper, there was a bit of a misunderstanding regarding our dataset. We originally thought we were going to be working with average review scores (ex. 4.6/5 or 9.2/10) as opposed to the rankings of the strains on Leafly (1 being the best). This posed an issue because just by looking at the dataset, it looks like the lowest ranking is in the 140-160 range. To put things in perspective, we have hundreds of thousands of strains in our dataset.



However, we ultimately decided to build the model anyway and see if we could make any decent predictions. Looking at the graph at the right, I'm sure you can see why we came to this conclusion. The MSE was around the 2000-3000 range, which means our error on average was about 50. With the rankings being 1-150 (roughly), that makes this a very poor model. So, in conclusion, it seems other factors are more important in a strain's popularity. It would've been productive to test this with other variables such as strain type or producer.

Reproduction of Results:

To begin, you will want to pull the dataset from the DoltHub repository.

To pull the dataset from the DoltHub repository and convert it to a local CSV file, these are the steps you will need to take (credit for figuring this out goes to our mentor, Joshua Ervin):

1) Install Dolt on your computer
    a. Run *sudo bash -c 'curl -L https://github.com/liquidata-inc/dolt/releases/latest/download/install.sh | bash'* **(MacOS/Linux)**
    b. **Windows** users select a .msi file from here: https://github.com/liquidata-inc/dolt/releases
2) Clone the repository locally by running *dolt clone Liquidata/cannabis-testing-wa* on the command line.
3) Change into that directory (Liquidata/cannabis-testing-wa) and run the following command at the command line: *dolt table export –file-type=csv tests test_file.csv*
    a. This will generate a test_file.csv in that directory.

When opening the Python script provided, you'll notice every function is already called in the main() method. There are comments within the method that separate the three different parts of it that perform different tasks. The first part simply loads in the data, drops NaN values from it, and produces a restricted dataset (cleans it up) by calling one of the functions. For this chunk of code to work exactly the same on your end, make sure the aforementioned .csv file is named "cannabis-testing-wa.csv" and exists in the same exact directory and folder of your python file.

The second part of the main method prints out the correlation statistics that allowed us to determine which variables are more significant in determining cannabis potency (and which ones have no noticeable effect). While we provide code that prints out the statistics we used in consideration (independent variables in

restricted dataset), you may also find out the correlation coefficient for any of the columns in the untouched dataset by calling:

print(correlation_coefficient(data, '<variable_name>') # Note: < > is a placeholder

The third and final part of the main method calls every function that produces either a data visualization or statistics for a particular machine learning model. To produce the four data visualizations that don't pertain to a machine learning and/or regression model, type:

thc_year_to_year(data) # Shows average THC content in cannabis for each year

average_max_thc_value(data) # Shows the overall distribution of THC content values in the dataset

inventory_type_visualization(data) # Shows average THC content for each cultivation method

chemotaxonomy_visualization(data) # Shows average THC content for each chemotype.

All these function calls will save a .png image of the visualization in the directory/folder you save this Python file in. The last two function calls you'll notice in the main method pertain to the machine learning models:

thc_content_predictor(res_data) # Runs a model to predict THC content in strains of cannabis.

leafly_review_predictor(data) # Runs a model to predict the Leafly review ranking of strains.

Both function calls output similar things. Each will print the test mean squared error of the model and a DataFrame that compares actual and predicted values. They will also have their own visualizations that also compare actual and predicted values and are both saved in your directory.

We made sure to make the process of reproducing our results rather simple by calling everything in main and providing comments that indicate which calls do what. Simply opening the Python file and hitting "Run" if its an IDE, or calling *python predicting_thc_main.py* in the Terminal, will print out all relevant statistics and store the relevant visualizations. It should be noted that calling all these functions at once will cause the run-time to be rather slow due to the large amount of data that it's sifting through (hundreds of thousands of rows in the dataset).

Work Plan & Evaluation:

Estimate: Both of us will meet in-person and work together in pairs for both writing the code and writing the report. There will be no need to divide responsibilities and work separately. However, with writing code, there'll be points where we are both testing different things with our model and so we'll make use of a version control system such as Git for the sake of efficiency. Regarding the time-frame of our work plan and specific due dates we set for ourselves for different portions of the project, refer to the time schedule below:

Date/Timeframe:	**Objective:**
By Friday, March 6	Complete the data-cleanup (removing unnecessary columns from the dataset, dealing with any NaN values, etc.) and finalize testing different variables for their statistical significance in the model so as to begin the process of coding the model (for the first two research questions). Expected Time: 3-4 hours.

By Monday, March 9	Complete both models over the weekend, record our data/observations, and save/finalize any visualizations. Come together on a conclusion for our project (discussing whether or not our questions were answered effectively, if at all). Begin work on the first draft of our written report. Expected Time: 5-6 hours.
By Wednesday, March 11	Complete the rough draft of our written report and begin on the revision process (our own revision/editing along with any necessary review from peers outside of this course). Expected Time: 5-6 hours.
By Friday, March 13	Finalize our written report and submit all of our work.

Evaluation: Our work plan estimates were not very accurate to say the least. There was quite the unforeseen and unfortunate circumstance that arose around the time we planned to begin this part of the project, that being COVID-19. Our work plan was naturally hindered due to the overall disarray in our lives. We did not meet up to work on our project as many times as we had hoped to, and we ended up doing most of Part 2 in one day as a direct result of that. The hours were reasonably accurate, however. We spent a few hours cleaning up our data and determining the proper statistical means through which we would analyze a variable's correlation and relevance in predicting THC levels and popularity. The facet of this part that took us the longest was the creation of our model(s). We had to do quite a bit of research and sifting-thru-documentation in order to figure out how to perform a multi-regression in Python and subsequently visualize it.

However, our estimates for the kind of steps we'll take in completing this part of the project were actually quite good. We had to play around with which models we wanted to use, how we wanted to refine our data set, how we wanted to plot our data and results, and much more. The work plan was effective in that it provided us with a "check-list", if you will, to follow throughout the process. Regarding the mention of a version control system, we ended up doing most of our work in a Jupyter Notebook to determine the specifics of our code, and then transferred that code to a Python file. The notebook was on Google Colab so we were able to push changes to each other whenever necessary.

Testing:

With this project, we created various machine learning models in order to predict a strain's THC content and its popularity in the cannabis community. Because we created machine learning models, most of our testing is already incorporated into the process of producing each model. We split the data into a train set and a test set, with the test set being our "testing method." Also, a form of testing we undertook with these models is just testing different methods (LinearRegression vs. DecisionTreeRegressor, for example) and determining which one most accurately predicted what we were looking for. We did a lot of testing initially to ensure that our dataset was cleaned-up and refined, and this was done in a Jupyter Notebook so we can see the effects of each function call on the dataset rather quickly. We simply printed out our dataset as a DataFrame and visually analyzed the changes that were made. To test that our machine learning models were performing as expected, we graphed their results and produced DataFrames with actual and predicted values to compare with the mean squared error (this is all included in the source code). We also ran some statistical tests on specific variables (t

and f-tests) to identify and confirm which ones had the greatest influence. Statsmodels has built-in regression functions that allow one to do this, with the "summary" function including all the necessary statistics for this particular project (we call and print this out in the provided code).

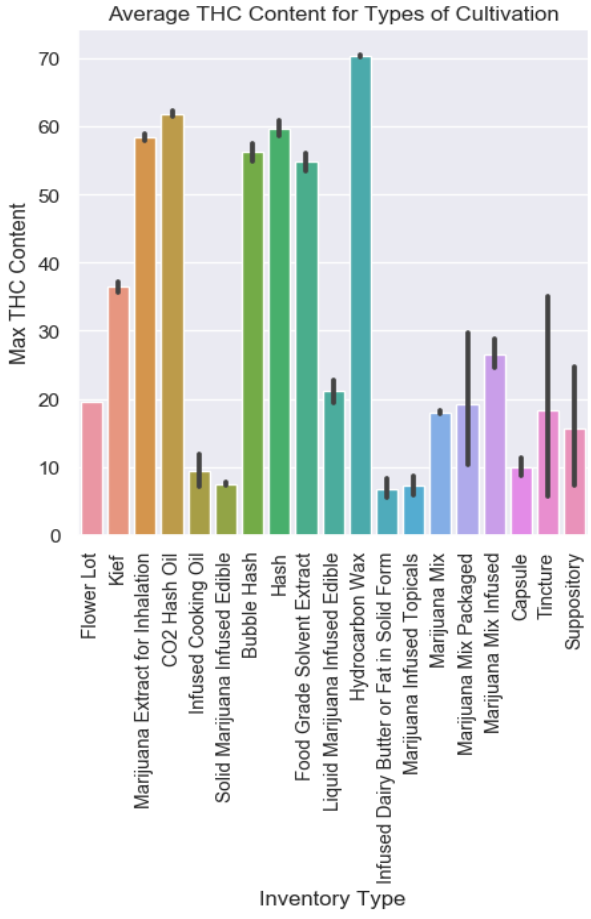
Our results are trustworthy because we show and acknowledge the error that lies in making such predictions. Our results do not assert that we can determine THC levels or a strain's popularity with near-perfect accuracy. We describe how it is possible to get an idea about a strain's THC content using our model, but we also mention that it is only moderately accurate at-best. We also provide a myriad of charts and graphs for showcase certain aspects of the dataset (Ex. Average THC level of the whole set shown in Appendix B), and we didn't necessarily need to. We built the functions that visualize the data in these ways in order to increase a reader's confidence in our ability to work with this particular dataset and extract meaningful information from it. Our results are based on accurate and recent data, and we specify/clean the data through very simple steps that we outline for the reader. The graphs we depict show some of the correlation that can be found between various variables, and our numbers prove that there exists some correlation, but we do well to reiterate that our models are not perfect.

Collaboration:

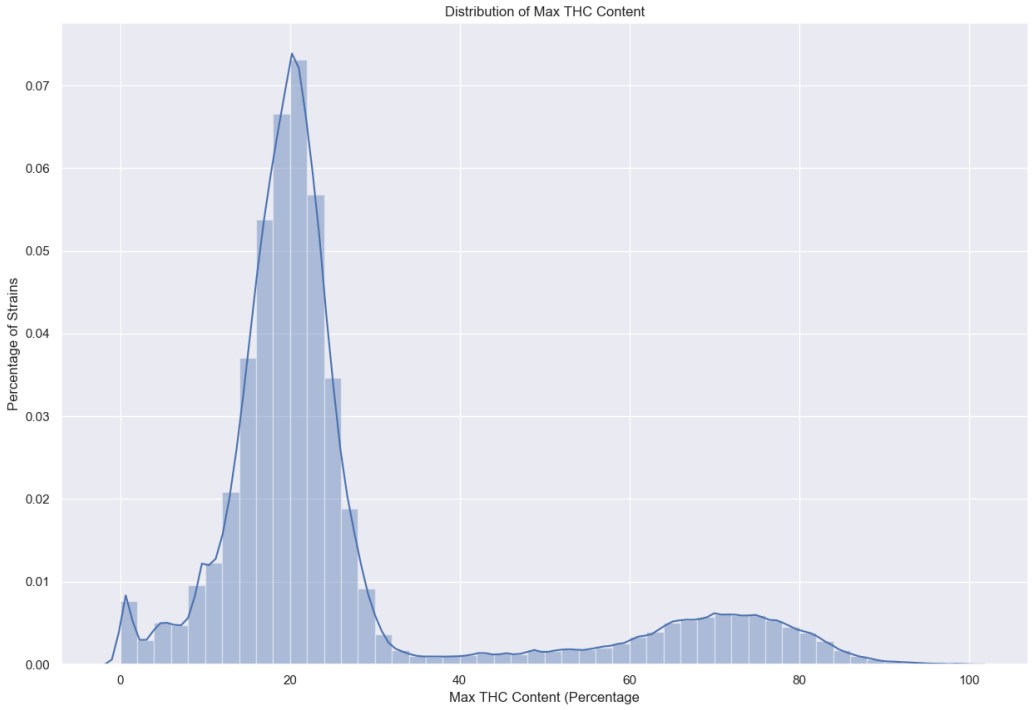
The only contributing members of this project are no other persons outside of course staff helped in the design and completion of this project. Copyright LLC ©

Appendix:

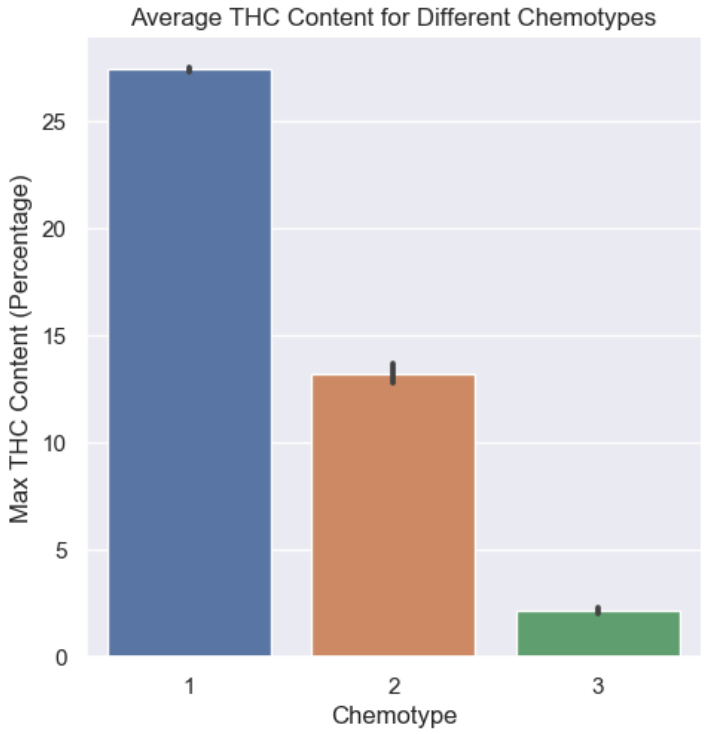
Appendix A: Average THC Content for Different Cultivation Methods (Inventory Types)



Appendix B: Distribution of THC Content in Dataset (All Observations Included)



Appendix C: Average THC Content for Each Chemotype (1-3)



Appendix D: Average THC Content in Cannabis Year-to-Year (2014-2017)