PREDICTING THC FROM BUD APOGEE

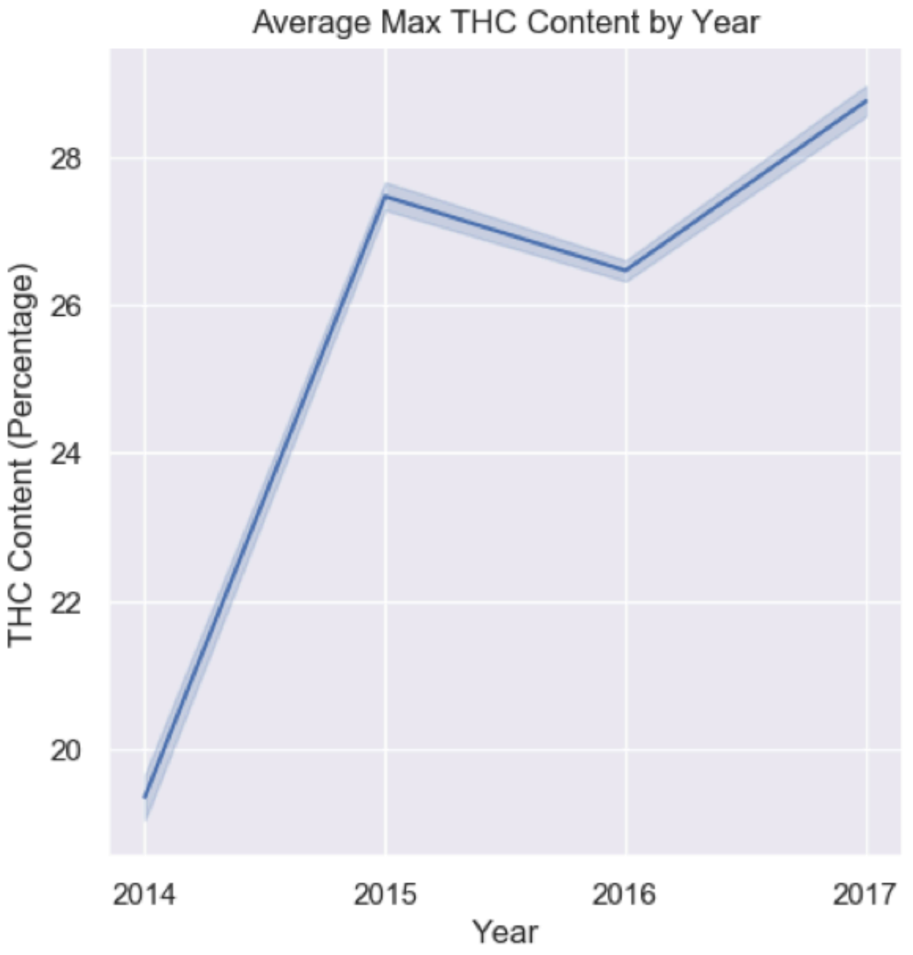
WINTER 2020

BACKGROUND & MOTIVATION

- Considerable increase in the potency of marijuana from 1995 to now [1]
  - About 300% increase in THC levels [1]

- Negative effects of cannabis primarily isolated and localized to THC [1]

- Concerning health risks for growing levels of THC [1]
  - Panic attacks, psychotic effects, paranoia,
  - Can produce massive vasoconstriction leading to decreased blood flow [1]



[1] Marijuana Investigations for Neuroscientific Discovery program at Harvard

BACKGROUND & MOTIVATION

- Users can be better informed about the weed they use
- Producers can understand the important variables in creating less/more potent marijuana so more likely to make a better product
- Increase efficiency in production due to less testing
- Understanding THC will help with laws regarding THC production & intake
- Dataset: over 200,000 laboratory measurements of cannabis products for legal sale in Washington state

RESEARCH QUESTIONS

Are some variables more important (stronger correlation) than others in determining THC content?

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

After predicting THC levels for specific strains, does understanding their respective attributes help us to predict their popularity in the marijuana community?

METHODOLOGY

- Clean and refine data set

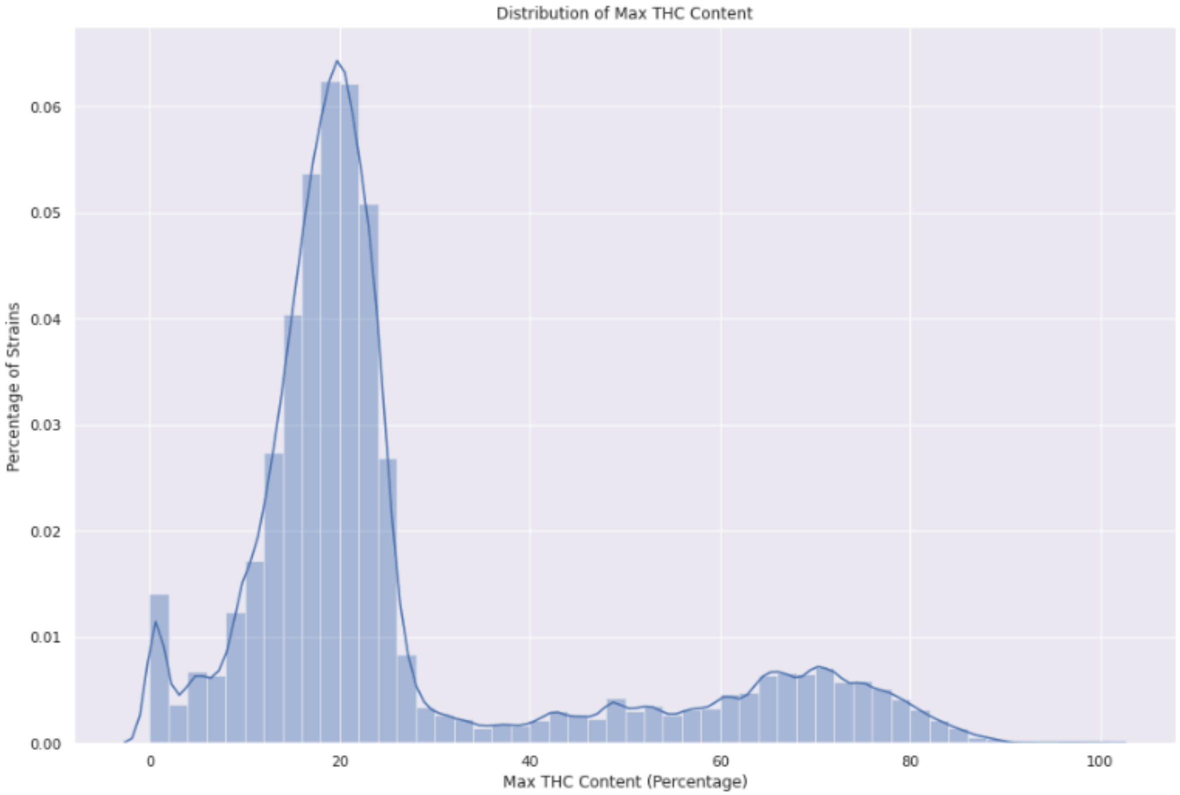
- Perform OLS regressions on the data to determine which variables impact THC levels the most
  - Determine from a returned correlation coefficient
  - Dependent variable: THC content
  - Independent variable: various columns previously deemed significant in data cleaning

- Create a decision tree regressor machine learning model to predict THC levels
  - Dependent variable: THC content
  - Test set and train set: 20-80% split
  - Determine using mean squared error

METHODOLOGY

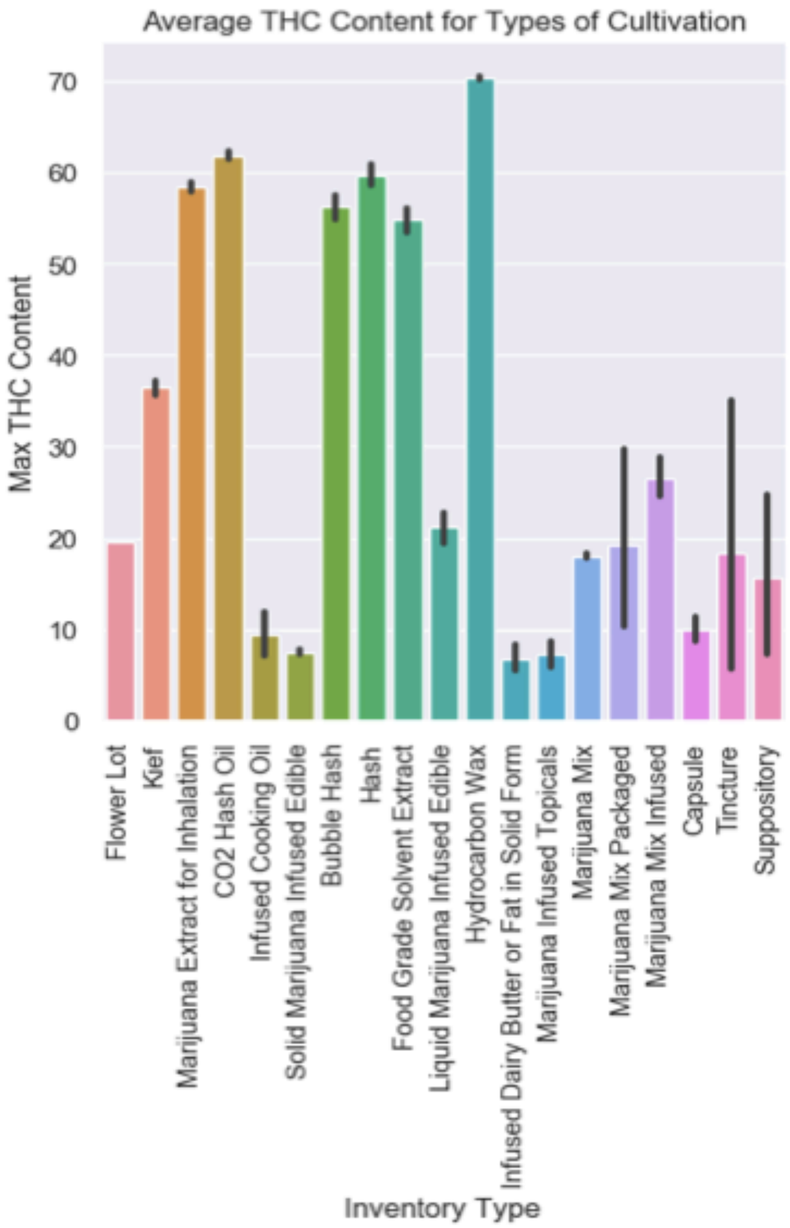
- Create a linear regression machine learning model to predict Leafly review ranking for a strain based on its THC content
  - Dependent variable: Leafly review ranking
  - Test set and train set: 20-80% split
  - Determine using mean squared error

- Plot all results found above appropriately using Scikit-Learn and Matplotlib



RESULTS

ARE SOME VARIABLES MORE IMPORTANT (STRONGER CORRELATION) THAN OTHERS IN DETERMINING THC CONTENT?


Average THC Content for Types of Cultivation
Max THC Content
0
10
20
30
40
50
60
70
Flower Lot
Kief
Marijuana Extract for Inhalation
CO2 Hash Oil
Infused Cooking Oil
Solid Marijuana Infused Edible
Bubble Hash
Hash
Food Grade Solvent Extract
Liquid Marijuana Infused Edible
Hydrocarbon Wax
Infused Dairy Butter or Fat in Solid Form
Marijuana Infused Topicals
Marijuana Mix
Marijuana Mix Packaged
Marijuana Mix Infused
Capsule
Tincture
Suppository
Inventory Type

- From the OLS technique, determined most significant variable is the cultivation methodology ('inventory_type' as its called in the dataset)

- Approximately 80% of THC content variation is explained

- The form of cannabis has a direct relationship with THC content

- Bar graph highlights different forms and their average THC content

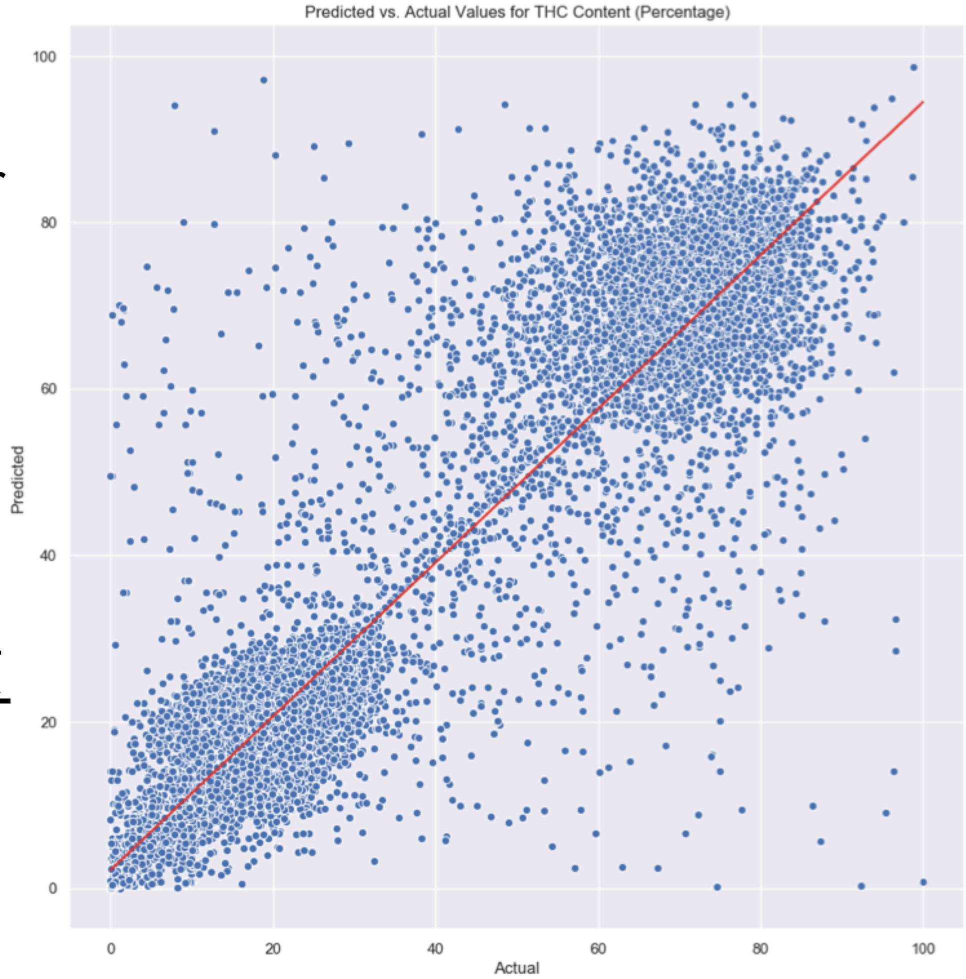
- Found that chemotaxonomy (chemical make-up of the plant) explains about 60% of the variation in THC content

- Both CBD level and Strain Type (sativa, indica, etc.) explain pretty much none of the variation

RESULTS

HOW ACCURATELY CAN WE PREDICT THE THC LEVEL IN A LEGALLY GROWN STRAIN OF CANNABIS?

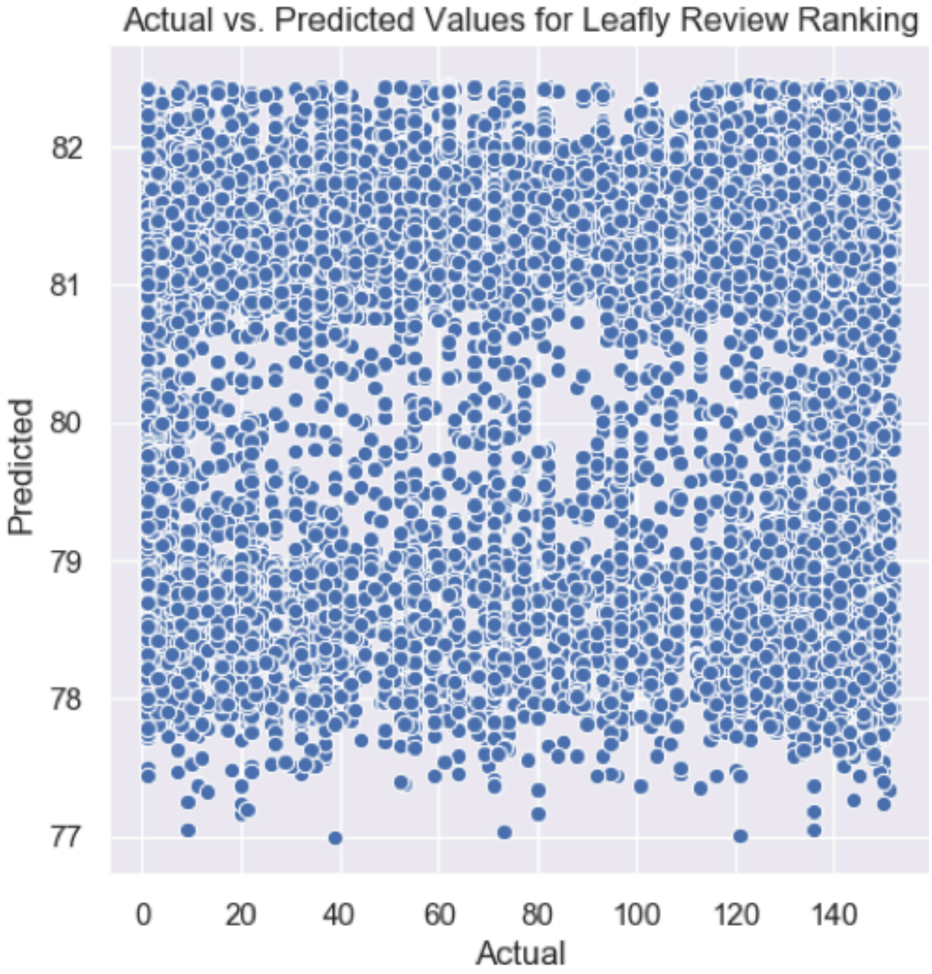
- We can predict moderately accurately
- Our machine learning model produces a mean squared error (MSE) of ~51
- This means that our error, on average, was roughly 7 when THC content is valued at a range of 0-100
- This isn't very good, but it's not bad either: model predicts near-perfectly about half the time
- Every independent variable in the restricted dataset was necessary to produce the best possible model



RESULTS

DO THC LEVELS HELP US TO PREDICT A STRAIN'S POPULARITY IN THE MARIJUANA COMMUNITY?

- We can not predict popularity in the marijuana community as represented by Leafly knowing a strain's THC content
- Originally thought data had average review scores as opposed to the rankings on Leafly so it made it very hard to create any sort of correlation
- Produces a mean squared error (MSE) in the range 2000-3000, 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



THINKING TO THE FUTURE

- Find more data sets about legally grown cannabis
- Do more research as to what variables really impact THC
- Perform better statistical analysis in determining important variables
- Create and use a more complex machine learning model to predict THC levels
- Create an app that allows users to simply enter a few descriptions and facts about their weed, to the best of their knowledge, and returns an estimated THC content based on what they inputted