Pete Schultz

INST 462

Mini Project: Exploratory Data Analysis

The domain I will be focusing on for this project is Biology, specifically marijuana plants. The question I would like to answer is: *What are the top 10 strains that are used most for breeding new strains? And among these top 10, are there any similarities or patterns that can be found among their children?*

I began this process by collecting the data myself from Leafly.com using web scraping tactics in R. The dataset was built on March 24, 2020 and the code can be found in leafly.R. Next to discover the top 10 strains that are used most for breeding, I decided to put all of the strains in a network. I accomplished this using pandas (python) and created nodes and edges files that I would then load in the network visualization tool Gephi. This is a directed network with each of the child strains pointing to their parent strains. I expanded the graph using Force Atlas with a repulsion strength of 10,000, color coded the nodes based on their type (hybrid, indica, or sativa), and sized them on degree. Last, I filtered the labels to appear only on the top 10 highest degree strains and vertically aligned them in the center for visual aid.

Now that I have gotten the top 10 parent strains of the population, I wanted to visualize them with their children on a scatterplot, specifically comparing %THC vs energy level. I achieved this by creating csv files for each strain using pandas (python) and loaded them into Tableau. All of the python code used for this project can be found in strains.ipynb or strains.html. To create these plots in Tableau, I placed THC on rows and Energy on columns, filtered out strains with missing values under either THC or Energy, color coded them by type, and shaped them by either child or parent. Ten individual plots were created along with 10 gif animated visualizations of the plots, which were made by exporting an image per added child to the plot and inputed all of them for each strain into the gif building website EzGif.com.

Finally with all of visualizations finished for the project, I put them all together on a website I created through Wix. The network visualization is found at the top of the page along with the top ten parent strains with their plots and list of children. Some interesting findings were that Haze and Jack Herer usually bred high energy strains, while Afghani bred low energy strains, and all of the top ten typically bred children between 16-20% THC. However, I must note that most of the children for each of the top ten strains did not appear on the plots as they had missing values, which can be found on the website in bold under children.