Understanding Relationships Between Northwestern U.S Plants

DATA 301 Final Project
Andy Do, Jenisa Nguyen, Thea Yang
16 March 2021

I. Introduction

Our project is focused on analyzing plants using data science and statistical techniques. We were interested in seeing if data could summarize and show relationships between plants and give us a broader understanding of common trends in plants and how they grow. To accomplish this, we used the Trefle REST API, which provides information data for over four hundred thousand plants. More specifically, the API provides data on many different plant fields like the locations of where species typically grow, the conditions required to grow, and the features of the species, such as the foliage type, height, and color. As will be further explained, our focus of analysis will be on plants mainly located in the Northwestern United States.

II. Data Collection and Analysis

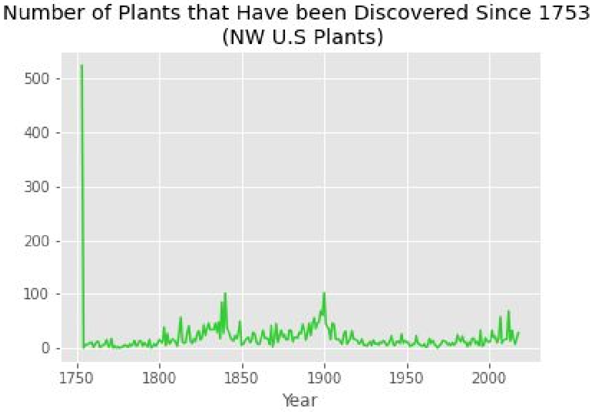
Data Collection and Cleaning

One of our main challenges with data collection and cleaning was the large API structure and vast number of information. This created a problem for us since the API had a limit of 120 requests per minute. Along with this, the request to obtain a list of plants did not include detailed information about the plant, and a subsequent request to the API would be needed in order to get more specific data. We decided to limit ourselves to just plants that were located in the North Western USA. The API has the location of plants stored using the ID from the World Geographical Scheme for Recording Plant Distributions (WGSRPD). Each request from the API would only produce twenty results so we used a for loop to iterate through all the plants located in NW USA. From the plants that we got, there were a few repeats so we first filtered out all repeated rows so that we would not be wasting time requesting information on the same plant. There were two columns that showed the link to more specific data about each plant. While trying to request each plant that was in our dataframe, we encountered a few errors where the link was not actually correct when used in the API. To try and fix this, we used a try/except block on both links and if neither of them worked we would just skip it and go on to the next plant. In order to save time, our group split up calling each individual plant into three groups and then combined them together at the end. Lastly, we filtered out all columns containing “common_names" since these columns just gave the name of the plant in different languages. This resulted in us obtaining a dataframe with detailed information about each plant that is located in North Western USA.

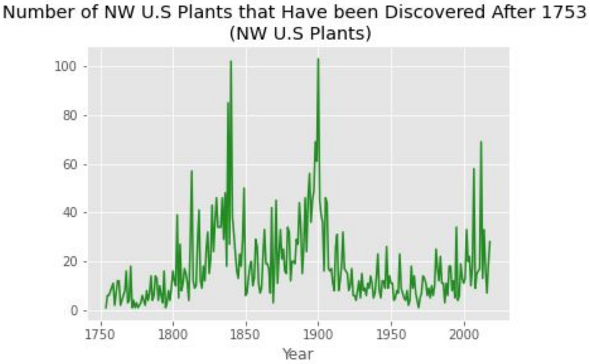
Data Exploration

After cleaning our data, we decided to explore the dataframe and the various information. As introduced before, we understood that our initial dataset included various different languages, which made it more difficult to analyze. That being said, we decided to use the Northwestern USA general location because of its proximity to us, as well as, the data allowing us to read in the data with ease. Overall, our dataset for Northwestern plants had 5373 rows of plant data, however many plants were missing data for certain variables, so you will see in our analysis that we often had to work with filtered data smaller than our initial overall dataset.

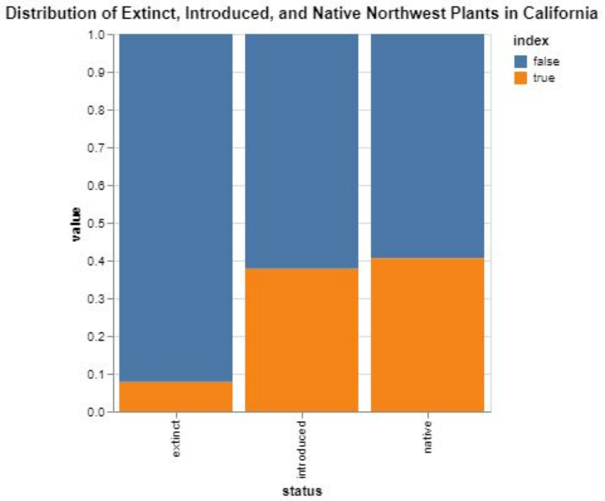
For the first set, we looked into the discovery dates of each plant. As shown in the graphs below, we saw the number of overall plants discovered was varied and had no major spikes, except around 1753. This spike is caused by the estimated date that several plants were discovered. As such, there is a larger spike here than in the other years.


Number of Plants that Have been Discovered Since 1753 (NW U.S Plants)

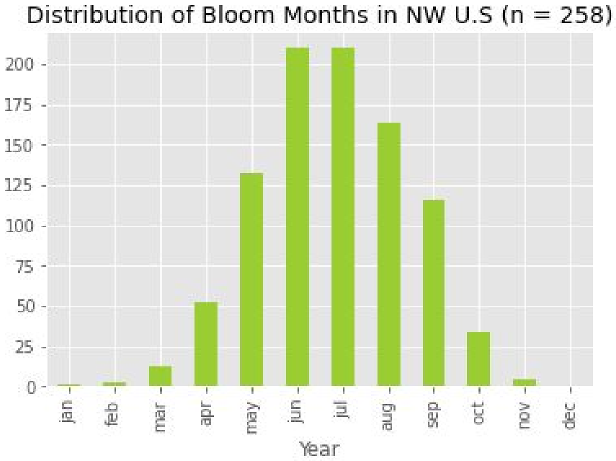
We also looked at the North Western plants that were discovered after 1753 to narrow down our search and allow for more analysis. This new graph demonstrates that there were a larger portion of plants discovered in the 1830s and 1840s. During this time, there was the Wilkes Expedition, which included botanists, who explored the west coast of North America. Additionally, in the 1890s, there were three other explorations with naturalists along the California coast. Moreover the third peak in the 2010s was likely the following of the rise of technology.


Number of NW U.S Plants that Have been Discovered After 1753 (NW U.S Plants)

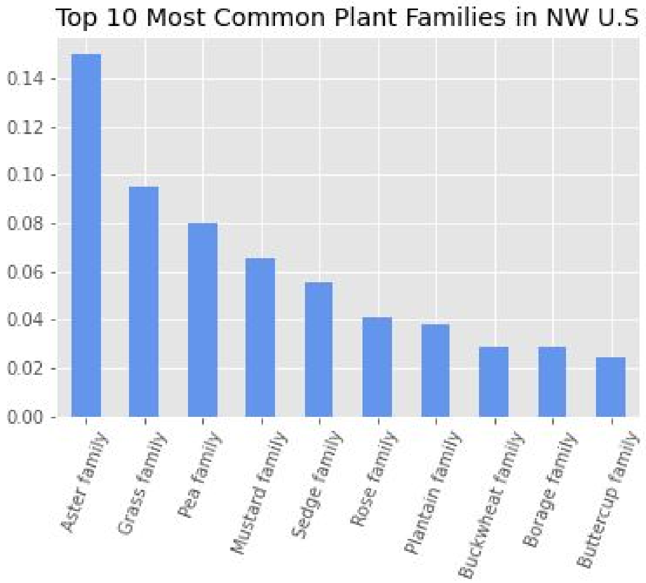
Next, we sought to check how many Northwestern plants were extinct, introduced, and native specifically in the state of California. This figure demonstrates that only 10% of Northwestern plants are extinct in California, while there are approximately 40% that were either introduced or native to the land. Although this does not demonstrate the overlapping distribution, we see an interesting observation that while there are few extinct Northwestern plants in California, most plants that are generally in the Northwest region are not introduced or native to California.



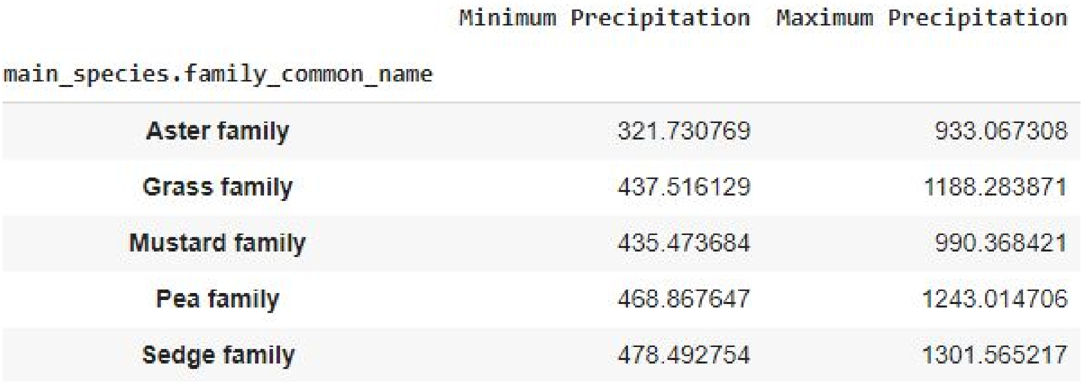
Going back to looking at Northwestern plants more broadly, we wanted to learn more about how the growth patterns of plants were distributed throughout the year. To do this, we looked at the distribution of months that plants were in bloom. However, because many plants were missing data for this field, we were only able to make a distribution using 258 Northwestern plants. In the figure below, it demonstrates how the summer months have the most number of blooms. This can be due to the amount of sunlight exposure that plants have in the summer and in effect allow for plants to gain more energy and resources. Similarly, we see how there are very few plants that bloom in Winter. While these observations may be fairly common sense, it is interesting to see it statistically explained by data. It's also interesting that while we typically assume Spring to be a season of growth, many Northwestern plants in this group don't start blooming until May.



For our next investigation, we first looked into the top 10 most common plant families in the Northwest US. In the figure below, approximately 16% of the eligible data came from the Aster family, and the Grass family followed behind making up approximately 10%. This graph was interesting to us, because we could get a rough understanding of common plants in the Northwest like pea, mustard, and buckwheat.

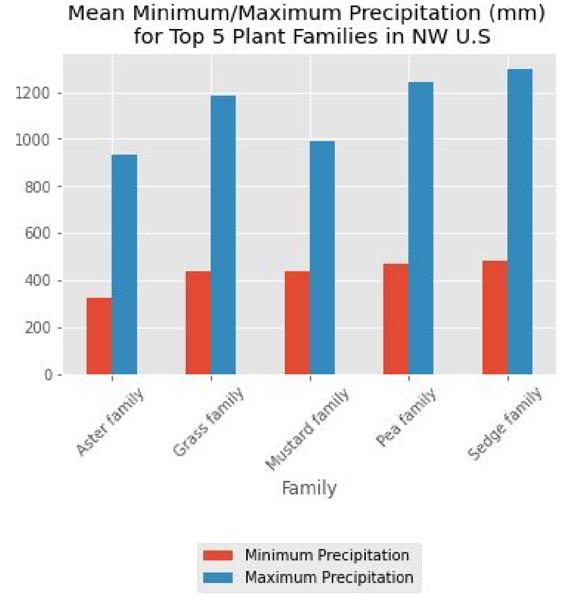


From here we were interested in learning more about plants relative to their families. More specifically, we wanted to know if plant families on average required different amounts of precipitation. To accomplish this, we narrowed down to the Top 5 most common plant families for ease of data access and calculated both their mean minimum and maximum required precipitation by family. Below is a table providing some statistics on these precipitation averages (in mm per year).

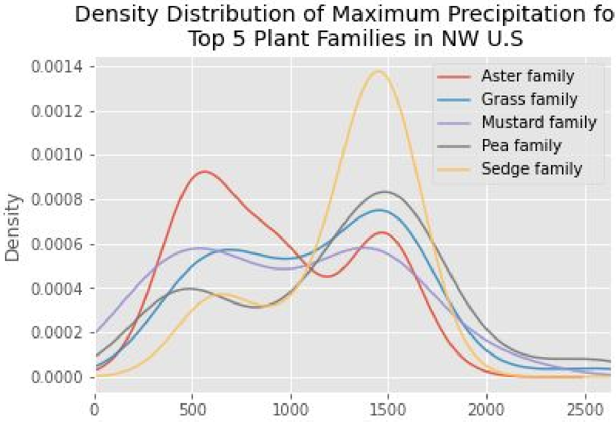


	Minimum Precipitation	Maximum Precipitation
main_species.family_common_name		
Aster family	321.730769	933.067308
Grass family	437.516129	1188.283871
Mustard family	435.473684	990.368421
Pea family	468.867647	1243.014706
Sedge family	478.492754	1301.565217

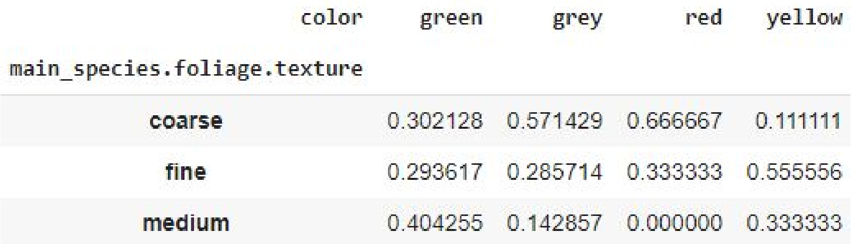
However, plotting these statistics gives us a much better visualization and way to understand the data. We plotted the minimum and maximum precipitation on the same graph, but note that it logically follows that the mean maximum precipitation for any family should be higher than the mean minimum required precipitation. However, this graph also allows us to compare between families. We see that while the mean minimum required amount of precipitation for the top 5 plant families are very similar, we see that there is more variation between the maximum amount of precipitation. The graph shows that plants in the mustard and Aster family on average tolerate less precipitation, whereas the Pea and Sedge families can tolerate a lot more precipitation on average, with the Sedge family on average having a maximum precipitation of around 1.3 meters! Moreover, this graph is interesting because it provides some insight on the relationship different types of plants have with water, and what plants need more or less water.



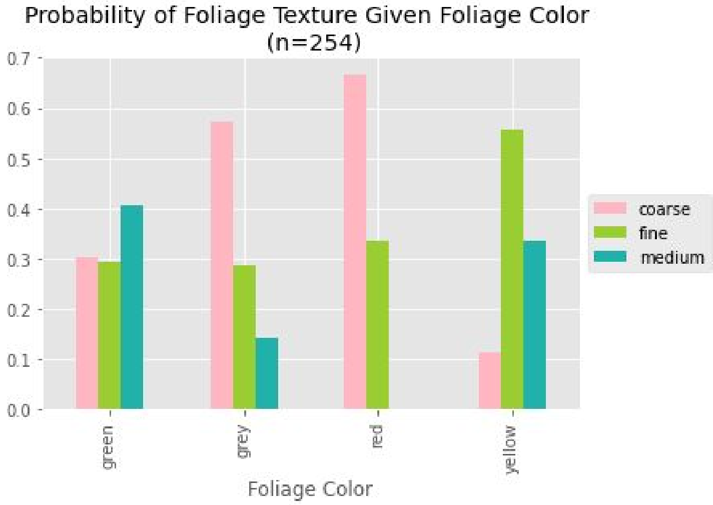
To go into even more depth in understanding this relationship plant families have with water and precipitation, we looked at the density distribution of maximum required precipitation for the top 5 plant families. This density distribution would help understand and illustrate the variation in maximum precipitation requirements for the different families, that the single statistic of mean can't show on its own. The graph below shows a very interesting result, mainly for the Aster and Sedge family. As shown previously, the Aster family had the lowest mean maximum precipitation among the top 5 families, and the Sedge family had the highest. This result is even more emphasized in the density distribution, where it shows the Aster and Sedge family having very distinct centers for their individual distributions. The Aster family is also the only family amongst the top 5 families that doesn't have most of their plant species in the 1 to 1.5 meter range. Alongside our previous observation that Aster family is the most common plant family in the Northwest, this provides interesting insight in relation to how much water and precipitation these plants need and the general environment and precipitation circumstances in the Northwest.



Finally, we investigated a topic that would lead into the main subject of our machine learning analysis. One of the plant fields given in the dataset describes the general foliage texture of a plant to be either “coarse”, “medium”, or “fine”. We wanted to see if there was a relationship between foliage texture and the foliage color of a plant. Many plants had missing data for either one of these fields or both, so the final filtered subset of the data included only 254 plants. We chose to look at the conditional probability of a plant having a certain foliage texture given the color of their foliage. Below is a table of the conditional probabilities and the corresponding graph. Some interesting observations include that given a plant being grey or red, the probability that the plant has a “coarse” texture is over 50%. On the other hand given a plant being yellow, the probability that the plant has a “medium” texture is over 50%. On the other hand, the probability of a plant being a certain texture given that the plant color is green is fairly evenly distributed. These results make sense considering that green foliage plants should be a lot more common, and thus plants with different colors of foliage might be skewed differently.



color	green	grey	red	yellow
main_species.foliage.texture				
coarse	0.302128	0.571429	0.666667	0.111111
fine	0.293617	0.285714	0.333333	0.555556
medium	0.404255	0.142857	0.000000	0.333333



Machine Learning

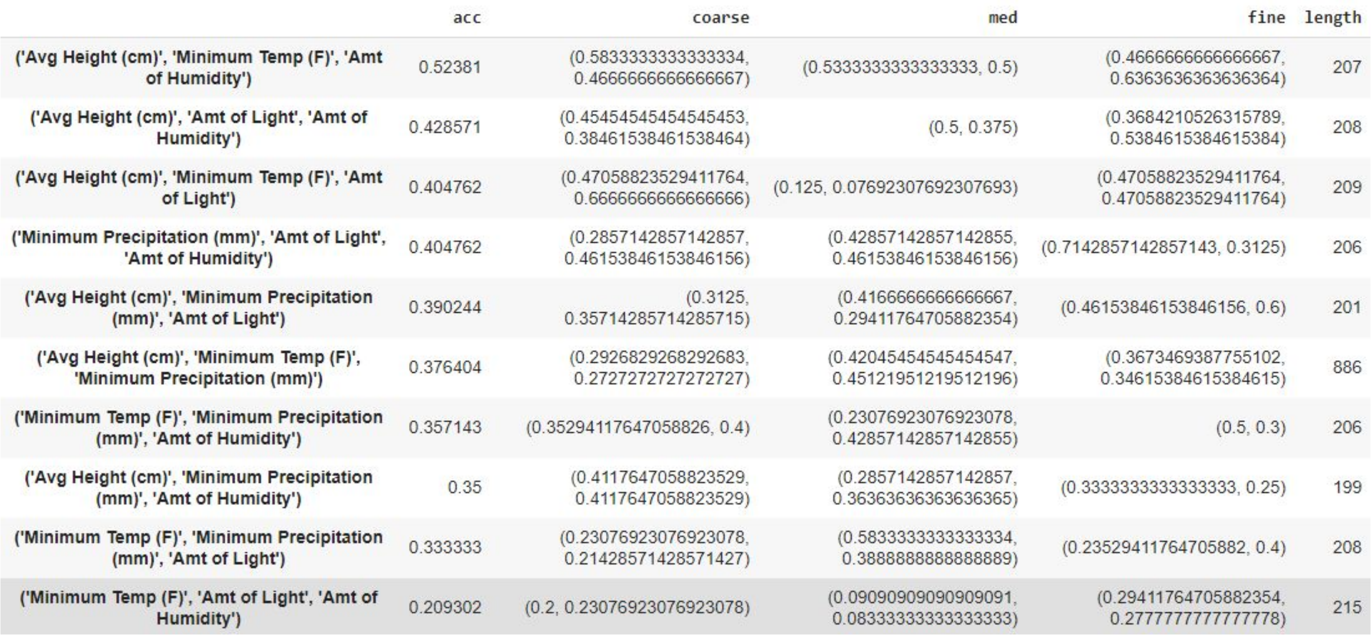
Previously we looked at the relationship between the foliage texture of plants and their foliage color, but we wanted to continue exploring the relationship between foliage texture with other plant fields. More specifically, we wanted to know if we could predict what kind of foliage texture a plant would have (coarse, medium, or fine) based on different features. In order to create this prediction model, we used a k-nearest neighbor classification model. A challenge we encountered when creating this classification model was choosing which variables to test in our model. A recurring issue with the Trefle dataset is that many plant fields were missing data, and there is no common pattern to what values are missing. Therefore when trying to select what variables to test, our dataset must be filtered to only contain plants with actual values for all the variables being tested. Starting with a dataset of 5229 observations, filtering this dataset drastically reduces the number of plants that we can use to build our model. Our solution to this challenge was to choose plant fields that had a high number of non-missing values for the goal of maximizing how big of a sample our classification model was built upon.

To build our model, we chose 5 different plant fields to test that all had over 800 non-missing values. These plant fields were the average height (cm), minimum tolerable temperature (°F), minimum required precipitation (mm), required amount of light (scale from 1-10)[1], and required amount of atmospheric humidity (scale from 1-10)[2]. These were all plant fields that we hypothesize may have an effect on foliage texture. To test which fields best predicted foliage texture, we tested different combinations of the variables and calculated the accuracy, precision, and recall for each model. More specifically, out of the 5 plant fields, we tested combinations of 3 variables at a time, for a total of 10 tested models. For each model we split the data into a training dataset for which the k-nearest classification model is built on and a test dataset, which we used to analyze the effectiveness of our model. Each time the dataset was randomly split 80-20, training and test data, respectively. Our k-nearest neighbors model used the value of 10 for k, and accuracy, prediction, and recall was calculated by comparing the predicted foliage texture from the test data with the true foliage texture from the test data.

Below are the results after testing the accuracy, precision, and recall for each of the 10 models. The "length" variable indicates the number of observations used to build the model. For each individual foliage texture, the first value is the precision, followed by the recall. The table is sorted by highest model accuracy.

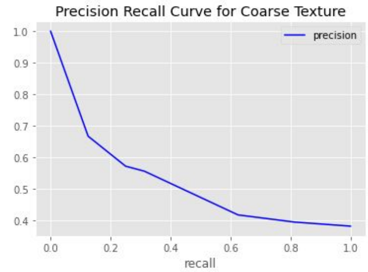
[1] Described by Trefle Documentation as: "Required amount of light, on a scale from 0 (no light, <= 10 lux) to 10 (very intensive insolation, >= 100 000 lux)"

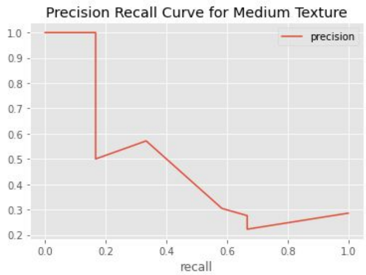
[2] Described by Trefle Documentation as "Required relative humidity in the air, on a scale from 0 (<=10%) to 10 (>= 90%)"

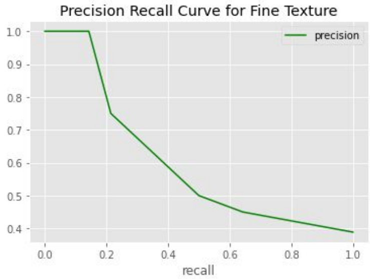


	acc	coarse	med	fine	length
('Avg Height (cm)', 'Minimum Temp (F)', 'Amt of Humidity')	0.52381	(0.5833333333333334, 0.4666666666666667)	(0.5333333333333333, 0.5)	(0.4666666666666667, 0.6363636363636364)	207
('Avg Height (cm)', 'Amt of Light', 'Amt of Humidity')	0.428571	(0.45454545454545453, 0.38461538461538464)	(0.5, 0.375)	(0.3684210526315789, 0.5384615384615384)	208
('Avg Height (cm)', 'Minimum Temp (F)', 'Amt of Light')	0.404762	(0.47058823529411764, 0.6666666666666666)	(0.125, 0.07692307692307693)	(0.47058823529411764, 0.47058823529411764)	209
('Minimum Precipitation (mm)', 'Amt of Light', 'Amt of Humidity')	0.404762	(0.2857142857142857, 0.46153846153846156)	(0.42857142857142855, 0.46153846153846156)	(0.7142857142857143, 0.3125)	206
('Avg Height (cm)', 'Minimum Precipitation (mm)', 'Amt of Light')	0.390244	(0.3125, 0.35714285714285715)	(0.4166666666666667, 0.29411764705882354)	(0.46153846153846156, 0.6)	201
('Avg Height (cm)', 'Minimum Temp (F)', 'Minimum Precipitation (mm)')	0.376404	(0.2926829268292683, 0.2727272727272727)	(0.42045454545454547, 0.45121951219512196)	(0.3673469387755102, 0.34615384615384615)	886
('Minimum Temp (F)', 'Minimum Precipitation (mm)', 'Amt of Humidity')	0.357143	(0.35294117647058826, 0.4)	(0.23076923076923078, 0.42857142857142855)	(0.5, 0.3)	206
('Avg Height (cm)', 'Minimum Precipitation (mm)', 'Amt of Humidity')	0.35	(0.4117647058823529, 0.4117647058823529)	(0.2857142857142857, 0.36363636363636365)	(0.3333333333333333, 0.25)	199
('Minimum Temp (F)', 'Minimum Precipitation (mm)', 'Amt of Light')	0.333333	(0.23076923076923078, 0.21428571428571427)	(0.5833333333333334, 0.3888888888888889)	(0.23529411764705882, 0.4)	208
('Minimum Temp (F)', 'Amt of Light', 'Amt of Humidity')	0.209302	(0.2, 0.23076923076923078)	(0.09090909090909091, 0.08333333333333333)	(0.29411764705882354, 0.2777777777777778)	215

The results of the above table indicate that the model with the highest accuracy was the model with the plant fields “Average Height”, “Minimum Temperature” and “Amount of Humidity”, where 52% of the predicted foliage textures matched the actual test foliage texture. Looking more specifically at the precision and recall for the three different textures, precision for “coarse” and “fine” textures were higher than recall, where the number of plants correctly predicted to be “coarse” or “fine” made up around half of the total number of plants predicted to be these textures. For “fine” texture, the precision was a little lower at 47%, but the recall was the highest, where 64% of plants correctly predicted the total number of plants that had a “fine” texture. Below are graphs illustrating the precision recall curves for this model for the three textures.







The three figures suggest that the default threshold for predicting the probability of texture (0.5) overall best balances the tradeoff between precision and recall, as for each graph we see that as we lower the threshold, recall increases, but precision also decreases.

III. Conclusion and Discussion

To investigate more about the relationship between plant foliage texture and many different growing conditions for the plant, we used a k-nearest neighbors classification model to see which conditions best predicted foliage texture. In the end, the best model we built was using the plant fields: average height (cm), minimum temperature (°F), and minimum required atmosphere humidity (scale 1-10). However the accuracy, precision, and recall were all fairly low even for our best model, with the accuracy being 52% and the precision and recall for each texture ranging from 40-60%. This suggests that the model can no better predict the correct class for foliage texture based on height, temperature, and humidity than a simple coin flip. While this leads us to conclude that our model was not that successful in predicting foliage texture, this also suggests that growing conditions like temperature, humidity, precipitation, and light, and features like average height have little influence on the foliage texture of a plant. This is an interesting result based on our initial assumptions that foliage texture might want to adapt to its environment and setting. The fact that the other 9 models we tested performed worse than this model indicates that perhaps foliage texture is influenced by factors external to the fields we tested from our model.

Looking at the data overall, many of our graphs and analysis can provide more statistical insight on the relationship between plants, whether it’s looking at the relationship between foliage texture and foliage color, or the relationship between plant family and required amounts of precipitation. We believe that analysis like this could be applicable to people and groups in the agriculture industry, and who could use these statistical results to guide and supplement their work. Part of data science and statistics is providing quantitative and data-based summaries and observations to make it easier to understand an intricate subject such as plants. Going forward, because collecting and cleaning the data from the API was challenging, it would be interesting given more computing power, to analyze plant data for a broader range of plants outside the Northwestern U.S region. Moreover, since the API was missing data for a lot of plant fields, we also believe that further refinement of the API might lead to more new and interesting data-driven discoveries.

IV. References

“A Global Plants API.” *Trefle*, trefle.io/.

“European and American Voyages of Scientific Exploration.” *Wikipedia*, Wikimedia Foundation, 19 Feb. 2021 ,en.wikipedia.org/wiki/European_and_American_voyages_of_scientific_exploration.

“List of Codes Used in the World Geographical Scheme for Recording Plant Distributions.” *Wikipedia*, Wikimedia Foundation, 23 Dec. 2020, en.wikipedia.org/wiki/List_of_codes_used_in_the_World_Geographical_Scheme_for_Recording_Plant_Distributions.