Locating the possible areas to discover cactus species using Apriori algorithm and visualizing the cacti spread across Arizona counties

Nagarjuna Battula
CISDE
Arizona State University
Tempe, AZ, USA
nbattul1@asu,edu

Himanshi Shrivastava
CISDE
Arizona State University
Tempe, AZ, USA
hshriva1@asu.edu

Priyanka Kumari
CISDE
Arizona State University
Tempe, AZ, USA
pkumari3@asu.edu

Rahul Arora
CISDE
Arizona State University
Tempe, AZ, USA
rarora16@asu.edu

Abhay Shrinivas
CISDE
Arizona State University
Tempe, AZ, USA
asarasw2@asu.edu

ABSTRACT

Cactus plants are one of the great surprises and paradoxes of nature. There are a lot of botanists putting tremendous efforts to know more about these plants. In this process, they are working really hard to locate and explore their characteristics by observing them over decades now. Our project aims to help them predict the occurrence of various cacti species in specific counties based on previous findings and generating rules using association rule mining. We also try to visualize the density variation of cacti species in the counties of Arizona which reveals interesting insights about their favorable weather conditions and habitat. These defensive spiny plants also produce magnificent blooms which are of research interest for botanists for various reasons that we cover in our project and our visualizations help them find fascinating patterns by going over their flowering seasonal changes. The project also helps us draw some inferences regarding the discoveries made in each county over a period of years by visualizing those trends. Thus, these findings and patterns try to tell us the story where gates for future research are always open.

KEYWORDS

Association rule mining, apriori algorithm, choropleth maps

INTRODUCTION

Plants are found all over the world. No matter what the weather conditions are, we find them. As the weather changes, their characteristics also change according to the place they are growing in. They adopt various habitats and live accordingly and thus contribute to ecological balance. It is easy for plants to grow in optimal weather conditions like good rainfall/ sufficient amount of water and enough nutrients in the soil. But what about the places where it doesn't rain? These dry and desert areas are very difficult conditions for the plants to grow in. However, as plants adapt to weather conditions in which they grow, we can also find plants in these desert conditions. Out of them, the popular ones are cacti and agaves. Even in the places where there is minimal amount of water and minimal amount of nutrition in the land, they grow really well. This has led to a research area where a lot of botanists are trying to find more about these plants.

From so many years, there are a lot of botanists putting tremendous efforts to know more about these plants. In this process, they are working really hard to locate these plants and know about their characteristics. They have been locating these plants and recording their observations from years. There are many botanical institutions working on collecting the data of plants in different parts of deserts in the world. One such institution is Desert Botanical Garden in Arizona. Their researchers have collected details of about 90000 specimens in the regions of the south west states of the US like Arizona and California and in northern Mexico [1].

From their research data, we have collected the data of the specimens that are recorded specifically in Arizona state. The dataset is acquired from the SEINET Portal Network and hence, it is in their format. All the observations about the plants and its location is in the *occurrences* data subset of the dataset, the images are in the *images* data subset of the dataset and the details about the identifications of the plants are found in the *identifications* data subset dataset. As the data is collected for over 90000 specimens the data was huge and many of the parts were unclean; just like a typical real-world dataset. We have made lots of changes and transformations in the data set and tried to answer some research questions that a botanist might find interesting and a readily available answer for those questions.

MOTIVATION

So, as we have seen, a lot of data is recorded by just one institution for a part of the world. This data is the result of the hard work and risks faced by all those botanists who were out there to collect details and observations on the plants. They have collected this data for a reason which is to make some interesting findings of the plants; know more about the plants growing in these dry regions. Arizona is one such region where we can learn a lot about these types of plants and our project is revolved around the counties of Arizona.

So, one of the most interesting question a botanist wants to know answer for is which of the plants are occurring in the same county in most of the counties of AZ. This is an important knowledge for botanists. Because in geography, there exist regions where only certain plants can grow, and they cannot be found in other regions because of various reasons.

The above question would lead to another research question. What species can be found in each county? And we are taking a unique approach to answer this question by considering the answer for previous research question. Let's put it in layman terms. Assume that we found two plants- A and B occurring together in almost all counties. Let us say that we discovered Plant A in a new county X. As plant A and plant B occurred together almost every time, we can say that there is also a chance for the plant B to be found in this county later in time. This is because the places in counties in which cacti grow, they all have similar weather conditions and soils.

Why the above research question can be useful to a botanist? As we have said earlier, a botanist takes lot of risks and does a lot of hard work. The temperatures in summer are usually above 100F in all the counties of AZ. And the cacti are one such species which grow easily in almost all the parts of AZ as AZ contains good weather conditions to support their life. This means that they can grow in empty lands with scorching temperatures or on hills which can sometimes be very steep and dangerous. Also, in many of the places where cactus grows, there are dangerous animals like snakes, lizards, scorpions, bees etc. wandering around. Not many of us would take up these challenges just to record the occurrence of a plant. And let's say the institute wants to find in which counties the plant B mentioned above grows. So instead of sending the recorders to all counties to find this plant, they can be sent to the counties possible to find this plant first. Thus, the search space to find plant B is limited here, making the botanists work easy.

Also, another research aspect a botanist might find interesting is that which counties have more of the cactus plants concentration and which counties have less concentration. More specifically, how many unique species of cacti grow in each county? With good techniques of preprocessing the data and visualization, we can easily show give an idea to the viewer how the cactus plants are spread across Arizona.

The cactus plants just like normal plants flower in various seasons. This gives rise to another research aspect where the botanists want to know in which months plants flower more and in which months plants flower less year-round. And also, if it flowers, which county flowers more in which month is also an interesting aspect.

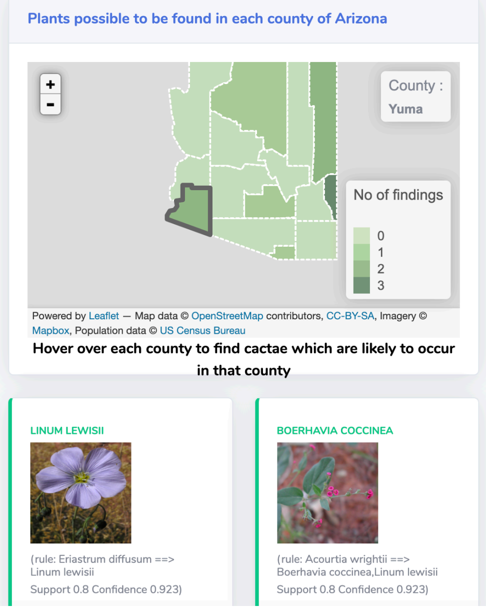
All these research questions can be answered by the biologists without any of the visualizations, but the thing is we are talking about over 50000 specimens in this dataset which belong to Arizona state. This takes a lot of time for botanists to find answers for these interesting questions without visualization. This summarizes the motivation behind our project.

VISUALIZATION DESIGN

Visualization 1: Finding possible plants to be found in each county

This county wise visualization aims to show the user the number of species that are not known to be there in the county but might exist along with their respective names. This visualization uses multiple cards. When the page is loaded, only one card is there. This card has county wise map of Arizona. The map is powered by Leaflet js [2]. All the counties are in various shade of green. As can be interpreted by the legend, higher intensity of green means that several species can likely occur in the county. As the intensity decreases, the count of species likely to be in the county also decreases. We go from a maximum of 3 species to 0 species. When you hover over a county, its boundaries are highlighted, and its name is displayed. Moreover, if likely count of species is one or more, new cards at the bottom are displayed. For every species that is likely to occur, there is a card. The card contains the name of the cactus and its image. Once, you hover off the county, these cards are not displayed anymore. Thus, user can always make out the likely areas where new locations of an already known plant can be found. The main advantage over here is the narrowing of the search zone.

We have used a choropleth map because we have very less number of colors to distinguish the number of findings across all the counties and a pleasant and active green color is used depicting the vegetation. Also, the rule which lead to the finding of the possible plant discovery in a county is also mentioned. Fig 1 shows one such activity of this visualization.


Plants possible to be found in each county of Arizona

Powered by Leaflet — Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox, Population data © US Census Bureau

Hover over each county to find cactae which are likely to occur in that county

LINUM LEWISII

(rule: Eriastrum diffusum ==> Linum lewisii
Support 0.8 Confidence 0.923)

BOERHAVIA COCCINEA

(rule: Acourtia wrightii ==> Boerhavia coccinea,Linum lewisii
Support 0.8 Confidence 0.923)

Fig 1. Finding possible plants to be found in each county

Visualization 2: Plants occurring that are found together in the same counties

The objective of this visualization is to make user effortlessly realize how presence of two different species is co related. We are making use of three cards for the task. The top left card on the screen, which serves as the main focal area, displays a county wise map of Arizona. This map is also powered by Leaflet js [2]. On hovering over a county its name is displayed. As desert is a large part of the state, the counties are initially showed in a shade of yellow. The card on the right shows a list of rules. These rules are in the form of Antecedent -> Consequent wherein both Antecedent and Consequent belong to the set of cacti species. The rule says that if there is occurrence of Antecedent, then we can be sure of occurrence of Consequent. Along every rule is a button that enables the user to visualize the corresponding rule. Once user clicks on the button, all the counties which satisfy this rule are highlighted in blue. Color of the other counties is not changed, thereby highlighting those specific counties to the user. The user then has an option of hovering over the counties and checking their names. On the bottom right of the screen, is another card, which display names of all the counties that follow the selected rule. In case no rule is visualized, the list is empty and county names are shown in this card. All of these cards ensure, user can quickly visualize the rules by checking how widespread it really is. The static image of this activity is shown in Fig 4.

Visualization 3 : Distribution of plants in AZ

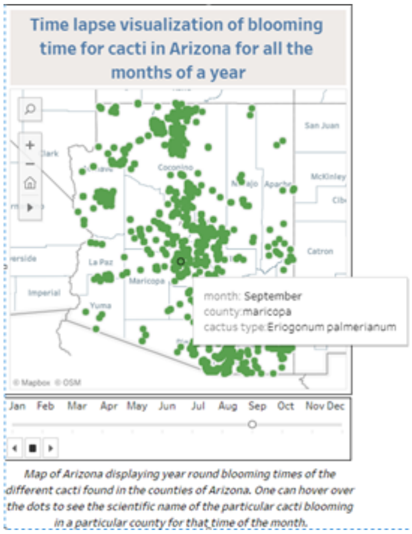
This visualization illustrates the difference in count of cactus plants over the various counties of Arizona. Bar graphs as show in the visualization clearly reveal the difference in the strength of cacti plant in all the counties. In just a moment one could realize how much difference there is in counties that have a lower number compared to the counties with greater numbers. It also provides a clear distinction of the counties in the middle against the counties that are towards either ends of the range. The color blue used for the depiction in this visualization matches with the theme of the website, thus adding on to the aesthetic. When the page is loaded, only a single card that has the bar graph is shown. When a user hovers over the bar, the total count in the county is shown along with a new card that pops up to the right. This card goes into a bit of detail and shows what the top 5 cacti are in the respective county. It also shows the number of these top 5 cacti, thereby helping one realize how the number is spread at the top. This data also enables a user to check if a species dominates over the others across the counties. The dynamic change in size of the card ensures no space is wasted if the text does not take up a lot of space. This ensures the text never appears too cluttered or too spacious, thus, adding a sense of uniformity. Fig 5 shows this visualization when Maricopa county is selected. This visualization was done using Chart js library[3].

Visualization 4: Blooming of plants in AZ round the year:

This visualization aims to provide information about the flowering pattern of different cacti across the Arizona counties over the span of the whole year. This was created to provide botanists an idea of the blooming season of various cacti found in Arizona which could lay them out with interesting information about the cactus biology. The tool used to achieve this visualization is the Tableau software.

For this visualization the most important data source is the **"**reproductiveCondition**" feature in the dataset**, which mentions the details of flowering/fruiting/vegetation/blooming of all the cactus species. We also have a month column which gives information about the growing months of cactus. So, a NULL value in the **"**reproductiveCondition**" feature** represents that the cactus does not bloom in the month associated with it. So, for a month all such records were not considered as reproducing months for the cacti. After filtering out this data, the cacti have to be grouped by month.

We used the Tableau software[4] to visualize this scenario by creating a flow map. Flow maps are a great way to display data change where something goes over time, in this case the blooming pattern. Firstly, to create the map we used the Latitude and Longitude coordinates for each data point in the path as rows and columns respectively. The county information is also used to add a layer of detail to aid form the map. Then we applied the filters described above for **"**reproductiveCondition**"** and month columns appropriately. To display data as points on the map we use the **"**scientificName**" feature in the dataset** which signifies each cactus species found in a county in Arizona. Now for grouping the data by month we used the pages feature in the Tableau software which provides the option to create multiple pages for a single sheet of visualization. This also allows us to bifurcate the dataset across the range of twelve months of the year, where we can select each month to see data corresponding just to that month. This feature provides a time slider with a forward and back option to change the slider range and a start/stop to play/pause visualization across all months at once. We also add additional information as a tooltip for each point in the map which are basically the month, county name and the cactus type. Fig 2. shows a snippet of the visualization for the month of September. From the visualization we can find that the flowering is usually less in winter months than in other months. This is a useful insight for a botanist.


Time lapse visualization of blooming time for cacti in Arizona for all the months of a year

month: September
county:maricopa
cactus type:Eriogonum palmerianum

Map of Arizona displaying year round blooming times of the different cacti found in the counties of Arizona. One can hover over the dots to see the scientific name of the particular cacti blooming in a particular county for that time of the month.

Fig 2. Blooming of flowers in September

Visualization 5: New Cacti discovered

This visualization is created to show the evolution of cactus discovery in the different counties of Arizona over several decades to effectively analyze the data geographically and over time. The map construction in this visualization (done using the Tableau software) is quite similar to the previous visualization for the blooming pattern but for data representation we use the concept of choropleth maps instead which are good to demonstrate ratio data much like what we need to display here: the changing densities of cactus discoveries each decade over all counties of Arizona. We use the count of scientificName grouped over each county for a decade, to aggregate to a number which eventually represents a shade in the color progression. Here we used the green color for symbolic reference to the color of a cactus. Hovering over a county shows the actual count of new species discovered in the selected decade. Below is a snippet of the cactus discovery trend for the year 2011-2018**.**

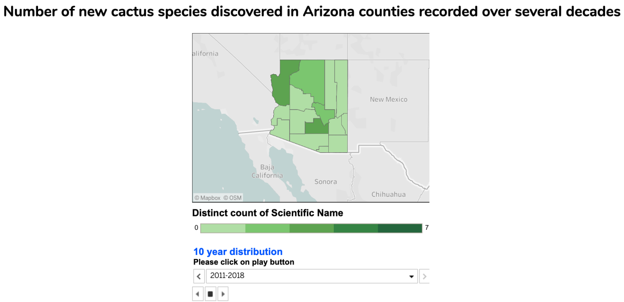
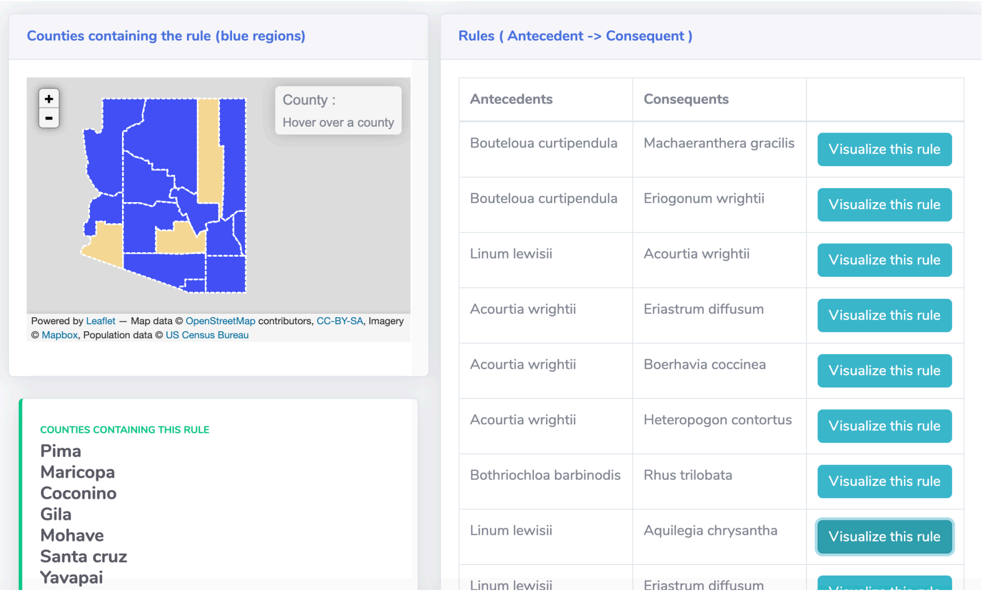

Number of new cactus species discovered in Arizona counties recorded over several decades
alifornia
New Mexico
Baja California
Sonora
Chihuahua
© Mapbox © OSM
Distinct count of Scientific Name
0
7
10 year distribution
Please click on play button
2011-2018

Fig 3. Visualizing the new cacti discovered in every 10 years in AZ

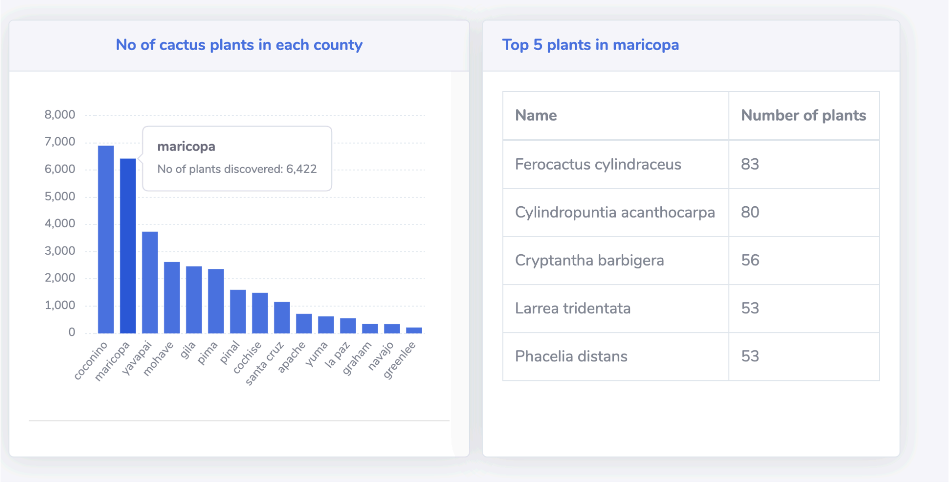

Counties containing the rule (blue regions)
+
-
County :
Hover over a county
Powered by Leaflet — Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox, Population data © US Census Bureau
Rules (Antecedent -> Consequent)

COUNTIES CONTAINING THIS RULE

Pima
Maricopa
Coconino
Gila
Mohave
Santa cruz
Yavapai

Antecedents	Consequents	
Bouteloua curtipendula	Machaeranthera gracilis	Visualize this rule
Bouteloua curtipendula	Eriogonum wrightii	Visualize this rule
Linum lewisii	Acourtia wrightii	Visualize this rule
Acourtia wrightii	Eriastrum diffusum	Visualize this rule
Acourtia wrightii	Boerhavia coccinea	Visualize this rule
Acourtia wrightii	Heteropogon contortus	Visualize this rule
Bothriochloa barbinodis	Rhus trilobata	Visualize this rule
Linum lewisii	Aquilegia chrysantha	Visualize this rule
Linum lewisii	Eriastrum diffusum	

Fig 4. Visualizing the rules



Top 5 plants in maricopa

Name	Number of plants
Ferocactus cylindraceus	83
Cylindropuntia acanthocarpa	80
Cryptantha barbigera	56
Larrea tridentata	53
Phacelia distans	53

Fig 5. Visualizing the spread of plants in Arizona

METHODOLOGY

As we are trying to answer some research questions in this project, we have used different visualizations to make it an easy process.

The intelligence aspect in this project is to give those counties of Arizona where the plants would be found later in time. The approach we have used to give those places is the Apriori algorithm and Association rule mining [5].

Apriori algorithm is used to generate frequent itemsets. In our case, plants occurring in different counties together are the frequent itemsets. The apriori algorithm for extracting frequent itemsets is as follows:

Consider plants A,B,C,D are there across Arizona counties and their counts are –

Plant	Count
A	2
B	5
C	3
D	4

Table 1: Occurrences of plants

Now there is a parameter called as support count. This is the number of occurrences of plants in counties. Let us set it as 3. This means that all the frequent itemsets will occur at least 3 times.

Now as you can see, A has only 2 occurrences. Thus, A can be pruned from the frequent itemsets.

So in the first step, the frequent item sets are:

Plant	Count
B	5
C	3
D	4

Table 2: Frequent itemsets in step 1

From these frequent itemsets, we generate more frequent itemsets by doing all the possible combinations of the itemsets in the table 2. The tentative output of step 2 (assumption) will look something like this:

Plant	Count
B	5
C	3
D	4
{B,C}	2
{B,D}	3

Table 3: Temporary frequent itemsets in step 2 of the Apriori algorithm

As you can see here, {B,C} doesn't meet the support threshold. So it is pruned. In this way the algorithm continues until there are no combinations left.

Now that we have generated frequent itemsets, we should generate rules for those frequent itemsets using another parameter called confidence.

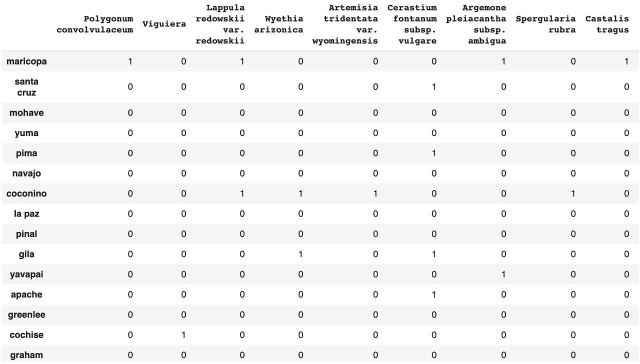
For itemsets X and Y, the confidence of the rule X ➔ Y =

support count of (X and Y) / support count of X.

Here X is called antecedent and Y is called consequent.

We set our confidence parameter as 60% and generated rules by trying all the combinations of frequent itemsets generated above in antecedents and consequents.

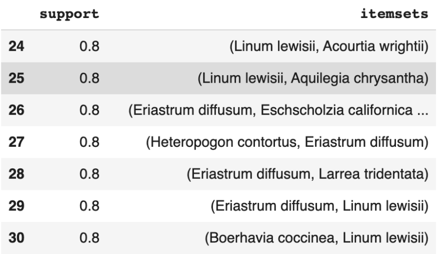
Coming to our scenario, We have taken the original dataset and transformed it to One hot encoding format as shown in the fig 6. below:



	Polygonum convolvulaceum	Viguiera	Lappula redowskii var. redowskii	Wyethia arizonica	Artemisia tridentata var. wyomingensis	Cerastium fontanum subsp. vulgare	Argemone pleiacantha subsp. ambigua	Spergularia rubra	Castalis tragus
maricopa	1	0	1	0	0	0	1	0	1
santa cruz	0	0	0	0	0	1	0	0	0
mohave	0	0	0	0	0	0	0	0	0
yuma	0	0	0	0	0	0	0	0	0
pima	0	0	0	0	0	1	0	0	0
navajo	0	0	0	0	0	0	0	0	0
coconino	0	0	1	1	1	0	0	1	0
la paz	0	0	0	0	0	0	0	0	0
pinal	0	0	0	0	0	0	0	0	0
gila	0	0	0	1	0	1	0	0	0
yavapai	0	0	0	0	0	0	1	0	0
apache	0	0	0	0	0	1	0	0	0
greenlee	0	0	0	0	0	0	0	0	0
cochise	0	1	0	0	0	0	0	0	0
graham	0	0	0	0	0	0	0	0	0

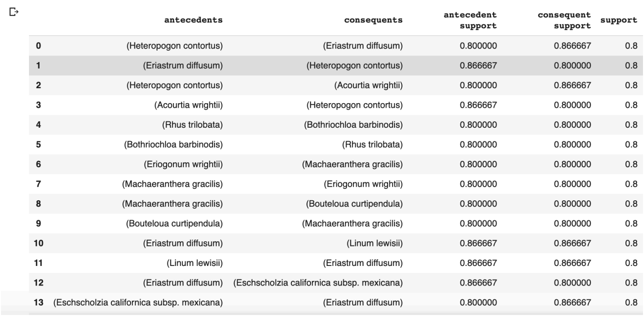
Fig 6. One hot encodings

In this data, if a species is present in a county, it is marked 1. Or else it is marked 0. For example, in the above figure, Viguiera plant occurs only in Cochise county so it is marked as 1 for Cochise county and 0 for all other counties. We have used this transformed data to generate frequent itemsets using apriori principle mentioned above as shown in fig 7.



	support	itemsets
24	0.8	(Linum lewisii, Acourtia wrightii)
25	0.8	(Linum lewisii, Aquilegia chrysantha)
26	0.8	(Eriastrum diffusum, Eschscholzia californica ...
27	0.8	(Heteropogon contortus, Eriastrum diffusum)
28	0.8	(Eriastrum diffusum, Larrea tridentata)
29	0.8	(Eriastrum diffusum, Linum lewisii)
30	0.8	(Boerhavia coccinea, Linum lewisii)

Fig 7. A snippet of the Frequent itemsets generated



	antecedents	consequents	antecedent support	consequent support	support
0	(Heteropogon contortus)	(Eriastrum diffusum)	0.800000	0.866667	0.8
1	(Eriastrum diffusum)	(Heteropogon contortus)	0.866667	0.800000	0.8
2	(Heteropogon contortus)	(Acourtia wrightii)	0.800000	0.866667	0.8
3	(Acourtia wrightii)	(Heteropogon contortus)	0.866667	0.800000	0.8
4	(Rhus trilobata)	(Bothriochloa barbinodis)	0.800000	0.800000	0.8
5	(Bothriochloa barbinodis)	(Rhus trilobata)	0.800000	0.800000	0.8
6	(Eriogonum wrightii)	(Machaeranthera gracilis)	0.800000	0.800000	0.8
7	(Machaeranthera gracilis)	(Eriogonum wrightii)	0.800000	0.800000	0.8
8	(Machaeranthera gracilis)	(Bouteloua curtipendula)	0.800000	0.800000	0.8
9	(Bouteloua curtipendula)	(Machaeranthera gracilis)	0.800000	0.800000	0.8
10	(Eriastrum diffusum)	(Linum lewisii)	0.866667	0.866667	0.8
11	(Linum lewisii)	(Eriastrum diffusum)	0.866667	0.866667	0.8
12	(Eriastrum diffusum)	(Eschscholzia californica subsp. mexicana)	0.866667	0.800000	0.8
13	(Eschscholzia californica subsp. mexicana)	(Eriastrum diffusum)	0.800000	0.866667	0.8

Fig 8. Association rules

From the frequent rules generated, we have generated association rules as shown in fig 8.

We used these rules in our visualization fig 5 to point out what are the plants that occur together in same counties. This will give botanist an idea for which species of cactus grow in same counties. We used these rules as basis to predict what plants might be found later in each counties of AZ. As we have talked about it earlier in the motivation section, let us take a rule plant set X ➔ plant set Y. Using this rule, we check all the counties containing all the plants in the antecedent. Then, we check all those consequents whether they are present in these counties are not. If one of the plant Z in consequent is not present in that county A, we can say that the Z is likely to be found in county A later in time.

A botanist can do all these calculations individually, but the thing is there are thousands of plants located in 15 counties of Arizona and these calculations give only the results in a format that requires a lot of time to keep track of with every finding in the result. But with the help of effective visualizations which can represent long matter in a short space, it becomes very easy for the botanist to identify in each county, how many plants are likely to be discovered.

Once the rules are generated and used the above logic, we have created a visualization as in fig 1. In the visualization, we have used the choropleth map of Arizona based on the results we got. This helps botanist identify how many plants can be found in each county just by looking at the map. As soon as the user hovers on a county, we have given the Name of the plant along with its image to identify the plant. So, the user will know precisely what plant could be possibly discovered in that county. Also, we have given the rule which lead to that finding as a botanist would want to know whether the finding that is generated is valid or not.

Also a botanist would want to know how the cactus plants are spread across AZ. For this purpose, we have used the visualization as shown in Fig 5. We have chosen a bar graph there keeping in mind that the visualization should be making good sense to both botanists and the general audience. Botanist would want an exact number sometimes and the might want to have an approximate idea of the number. In bar graph we have visualized it so that when you hover over a bar which represents a county, you get the exact number of plants. And when you see the visualization you get an approximate number. Also, we have included the feature of giving top 5 plants found in the county when you hover on a bar which represents a county.

Cactus flowers are one of the great surprises and paradoxes of nature. One must wonder how such spiny plants can possibly produce these magnificent blooms. It's like the instinct for reproduction competing with the need for survival. A botanist will definitely be interested in knowing which months the cacti flower all across Arizona for various reasons - when and where to plant a particular cactus for blooming? whether or not a cactus will bloom? when is the major blooming season? what weather conditions harbor the flowering of a cactus?

Our visualization in Fig 2 provides a solution for this by furnishing a time lapse visualization of the flowering pattern of various cactus species over different counties of Arizona by which a botanist can recognize intriguing facts. For example, if one analyzes our visualization, they can see patterns like the cactus flowering less in the winter months than the rest of the months. One can also realize that the center of Arizona mainly the Maricopa, Yavapai, Pinal and Gila counties have cactus species blooming all year long. Also, Navajo and Apache counties have the least blooming overall amongst all other counties. Such facts are quite interesting for botanists to develop certitude and do predictions about their research like what kind of land, weather conditions and time of the year are favorable for a particular cactus or species. Also, one of the great misfortunes in this world is to lose a flora species before we can understand its true worth. There are cases of them being lost faster than they are explored. This visualization can also be used to understand the cactus species which are low on reproduction rate or basically flower less and might require careful attention for preservation based on their potential for providing wildlife habitat or other natural desiderata.

People generally think Arizona to be barren land and a desert. We try to change this perception by showcasing how varied species of Cacti occur together in different counties. For several years, various researchers and botanists have been travelling and trying to find different species in the Southwest part of the US. We visualize their success stories where in a period of ten years various researchers found so many different species in different counties. We try to picture how dense a county is in varied species of cactus in Arizona. We also try to help botanists figure out where the maximum population of cacti is and where he could possibly look for more species.

In our visualization as shown in fig 3, we try to answer the question - Where the most distinct species were discovered in a time period. This could give two conclusions. In some counties, many distinct species were recorded, we can say it could be possible due to easily accessible terrains in that area, or that area being close human civilization. In the counties where fewer discoveries were made, it could imply that either the weather conditions doesn't suit much cactus growth and their reproduction or due to difficult terrains it was difficult to explore that area by foot. Thus, we could use modern day equipment like drone photography to study the terrain, find new species or study the blooming season of the cacti in that area. Thus, with a visualization which shows a time lapse of discoveries of cactus species made in the period of 1800s to 2018 can help us draw many inferences.

Evaluation Plan

In every project, an evaluation plan helps to track the progress of the project and the continuous changes required to improve the model. In our project, we have made a visualization model to showcase the rich flora in Arizona state and tried to predict which plants could occur in different counties. The evaluation plan tries to answer a few questions[6] which helps us track our progress.

Evaluation Goal

a. What does this evaluation strive to achieve?

This evaluation helps us keep track of the progress in the model, how the visualization would help people expand their knowledge on flora of Arizona state, USA and also familiarize us with some important cacti characteristics.

b. What is the purpose and use of this evaluation?

The purpose of the plan is to maintain the progress of the project, assess the data used for the model and parameters which could be used to enhance the visualization.

Why is the project needed?

This project helps in understanding the rich flora or Arizona, which cacti can occur together and help botanists in limiting their search space of finding species in the counties of Arizona. Also, it is not physically possible for botanists to explore all the flora and even if it would be, our model can help lessen that need. We try to make a model which could predict which species can possibly occur together in counties of Arizona.

Who is the target audience for this project?

This project aims at researchers and botanists who are studying the Agavaceae and Cactaceae Species in the Southwest part of the US, but we mainly concentrate on Arizona counties. However, the model and visualization are user friendly for the general public to understand as well and give them a basic idea about the characteristics of Agavaceae and Cactaceae in Arizona.

What are the project objectives?

This project aims to solve the ongoing questions in the field of Agavaceae and Cactaceae Species study. This project aims to predict the occurrence of species in a specific county based on some ML algorithms. This project also aims to reflect on the blooming conditions of the Cacti in Arizona state. We also aim at finding the top researchers of the area in a period of 10 years from 1800s to 2018.

What resources are available to the project ?

In this project, we have used the dataset of Desert Botanical Garden Herbarium, with emphasis on plants of Southwest, USA and northern Mexico. The project mainly focuses on the Arizona counties and which is almost 70% of the whole dataset. We have used Chart js library for our visualization, Tableau for interactive visualization and charts

What activities are being undertaken to achieve the outcomes?

We have continuous enhancements in the model and visualization, which focuses on how changes should be made according to the changes decided by the team members. We try to enhance the project to maintain aesthetics uniformly and keep in mind our color-blind audience as well.

What outcomes are produced from the activities?

We try to create a visualization model which is based in the dataset of Desert Botanical Garden Herbarium and focuses on botanists and general audience who are interested to know about the Agavaceae and Cactaceae Species. We have created bar charts and some interactive visualization which makes our model easy to understand. We have used Choropleth graphs to give a better idea of location in the visualization.

What are the program’s intended outcomes ?

This project aims to help botanists in reducing their search space by predicting the possible occurrence of Cacti species in the counties where it hasn't been spotted. It prevents researchers from going into dangerous terrains and exploring difficult habitats.

This project could give the general audience an insight about the species found in Arizona and which places they can visit to enjoy those blooms in which time of the year. We found that generally cacti bloom in spring and summers from March to October.

Evaluation Design

Resources Used: Dataset of Desert Botanical Garden Herbarium which includes data from the Southwest part of the USA and some parts of Mexico. However, we concentrate mainly on Arizona counties. To visualize the interactive visualization and bar charts and Choropleth graphs - Chartjs, amcharts4, leaflet and Tableau.

Data used: We have modified some data according to our requirements and used it to generate graphs and choropleth maps. We have Decimal latitudes and longitudes from the dataset and evaluate whether they are giving us appropriate results using actual longitude and latitude coordinates of that place. We also used months data for showcasing the reproducing time of the cacti species and which time of the year the flowers or fruits the most. We evaluated this with the overall cacti blooming season which helped in verifying our results. We also visualized the data of various researchers working in this area and which cacti were discovered in which county in what period of time. We found which researcher made the greatest number of discoveries of Cacti species and verified our results from the Wikipedia.

We visualized all these results in our project using Bar charts and Choropleth maps.

The bar chart Distribution of plants in AZ shows the max numbers of different cacti species found in each county which is simple to understand.

Choropleth Graphs: We try to predict the occurrence of the Cacti species in various counties of Arizona which could be easily seen in the choropleth graph of :Plants possible to be found in each county of Arizona in which if we hover over each county to find cactus which are likely to occur in that county, we can see which species occur in that county.

Plants Found Together: In this visualization, we see which rules are used in which countries. If we hover over the rules, which country uses this rule should be displayed. This gives an easy understanding of prediction of the species which could possibly occur in the specified county.

Blooming of plants in AZ round the year: helps one to understand when the cactus flowers the most and in which season. This gives the general public an idea about where they can visit for a beautiful flowering season of Cacti species in Arizona and during what time of the year. The researchers can study trends in change of the blooming patterns of the cacti species and which cacti flower instead of producing fruits. New cacti discovered: helps one understand which species were discovered when and where and by which researchers. It tried to give an overall idea about the counties where most of the species were found. It also tries to focus on the time period when the most discoveries were made.

Evaluation Analysis

Analysis of data: We have analyzed the patterns and rules in our data set using Machine learning algorithm Association Rule mining to get frequent itemset and derived rules which helped in predicting the occurrence of Cacti in different counties. We have used python to clean our data and apply the algorithm to the dataset. We also analyzed the graphs using Tableau which helped us in better understanding the pattern in the blooming seasons and discoveries made in the counties.

Discussion and future work

In all our visualization we have incorporated the basic visualization principles by Dr. Edward Tufte. “The four design principles used are: Visual integrity, Graphical Excellence, Aesthetic Elegance, Maximizing the data-ink ratio”[7].

There’s a plethora of ‘interesting’ details for researchers to observe. The plants that occur in the same county in most of the counties of AZ is an interesting detail (since it attempts to establish a relation between the region and the plants that grow there) that has been discussed.

Drone.

Another interesting detail that has been discussed is the species that can be found in each county (since it attempts to establish a relation between the region and the species).

Yet another interesting research aspect is the counties that have more of the cactus plants concentration and which counties have less concentration since it narrows down the possibility of finding cactus in certain areas.

There is a scope for a few extensions to our visualizations. More interesting rules can be generated based on the climatic conditions which can be extracted from latitude and longitude. We can visualize the rules (generated using Apriori algorithm) based on the occurrence of the species based on the temperature.

As the habitat column in the dataset has a description of the land texture, we can apply text mining to retrieve the exact texture and can narrow the search space further down using climatic conditions.

Another possibility is that the input can be of a cactus plant. We can identify its species and give a possible location in which it can be discovered.

REFERENCES

[1] "SEINet",[Online].Available: http://swbiodiversity.org/seinet/collections/misc/collprofiles.php?collid=5 [Accessed: 12-Nov-2019]

[2] "Leaflet",[Online]. Available: https://leafletjs.com/reference-1.6.0.html [Accessed: 12-Nov-2019]

[3] "Chart.js",[Online]. Available: https://www.chartjs.org/docs/latest/ [Accessed: 12-Nov-2019]

[4] "Tableau public",[Online]. Available: https://public.tableau.com/en-us/s/resources [Accessed: 12-Nov-2019]

[5] R. Srikant, “Fast algorithms for mining association rules and sequential patterns,” UNIVERSITY OF WISCONSIN”, 1996.

[6] “Evaluation Plan Template”, Centers for Disease Control and Prevention [Online].Available:https://www.cdc.gov/tb/programs/Evaluation/Guide/PDF/Evaluation_plan_template.pdf [Accessed: 12-Nov-2019].

[7] Edward Tufte, “Guidelines for Good Visual Information Representations”, Interaction Design Foundation, 2 Sep-2006. [Online]. Available: https://www.interaction-design.org/literature/article/guidelines-for-good-visual-information-representations [Accessed: 12-Nov-2019].