Mecklenburg County
Extension Master Gardener (EMG) Program

Group 5
Parag Saxena
School of Data Science, University of North Carolina at Charlotte

Introduction

North Carolina Cooperative Extension brings informal education and research-based resources to communities through the state's land-grant universities—North Carolina State University (NC State) and North Carolina A&T State University (NC A&T)—as well as county and tribal governments. Mecklenburg County Extension Master Gardener program offers several key services such as helping and educating people with common gardening topics and problems and providing resources to gardeners to achieve success. These programs include workshops, classes, and outreach activities. The program also provides online services, such as the horticultural help desk and training for volunteers. Volunteers must attend 40 hours of training, an exam, and an internship. Additionally, volunteers must attend ten hours of training and complete twenty hours of volunteer work each year to keep their certification. This program's main challenges are its reliance on volunteers and donations, the lack of awareness of the program among North Carolina residents, and the extensive training rules that require a significant time commitment.

There are also several non-profit organizations offering similar services, such as the American Horticultural Society, the Charlotte Garden Club Council, and the 100 Gardens Club. The Master Gardener Program does have several competitive advantages over their competitors, however. The program's affiliation with NC State and NC A&T gives it access to free horticultural research and knowledge bases from reputable sources, as well as non-horticultural resources from

NC extension. The volunteers of their organization offer a variety of gardening services. Their extension agents are university field faculty who provide high-tech, high-touch expertise to meet the unique needs of diverse clients. Currently, The Extension Master Gardener program depends on volunteers and donations to carry out its work.

The goal of this project is to provide the program manager with an up-to-date understanding of the program's current state, as well as recommended ways to improve it in the future. The objectives include anticipating the most common gardener concerns at a particular time or month, identifying the most relevant types of questions, and making them available to program managers. This provides greater visibility across the organization and provides stakeholders with better insights. This is ultimately used to estimate the number of volunteer hours spent in the field and the average increase in donations.

Research and Analysis

We used three different data sources for our analysis. The first is the question and response data that was transferred from email to a spreadsheet, containing 7000 records. This data source contained the date, zip code, and questions and answers from the horticultural help desk from 2012-2022. The second data source was the volunteer data that was in excel, with 2144 with attributes such as Last name, date, volunteer hours, travel hours, program during 2018-2022. The last data source was the donation information in an excel sheet of 31 records with attributes such as date and donation amount during 2018-2022. Notably, the email message data comes in two sets in our analysis. Gmail Dataset 1 includes a full population of email messages communicated between residents and EMGVs dating back to 10 years, and Gmail Dataset 2, which represents a smaller subset of above and includes some additional attributes. This 2nd set was originally provided in spreadsheet format.

Question and Responses-Gmail Dataset 1

One key challenge for this data set was a large file (2GB size) that had a proprietary format (.MBOX). Before using ETL steps, we finally used a commercial tool to convert this data to spreadsheet format (flat file) using the MBOX to CSV Convertor tool. The raw file was successfully converted to a standard flat file format (CVS, 500MB size) that could then be fed into our SISS term analysis package.

Then, we developed a SSIS package to extract and score the most frequently used English terms from the flat file. In this step, we explored different ways of analyzing the flat file by running the SSIS package multiple times changing the following key parameters: Term type, Score type, Frequency threshold, Max length of terms. After obtaining the output from initial ETL steps, the scored terms were reviewed to additional processing: Terms not related to Gardening were removed from results; Top 125 gardening terms were reviewed to identify common gardening themes or Categories (Appendix A figures a, b, &c). This analysis helped convert unstructured data into structured data to provide insight on rankings of gardening Categories, as well as rankings of key gardening Terms (across both full messages and Categories).

The terms, categories and their rankings were then used to build a dashboard in tableau (Appendix A figure 1(d)). In Appendix A figure 1(d), the bar chart on the left shows 5 identified categories from the least frequent to the most. The heatmap displays the relative frequency for each word and is separated into different categories by color variation. At the bottom of the dashboard, we show the most frequent terms across all categories.

Questions and Responses - Gmail Dataset 2

Initially, the names for fields were changed to meaningful names, instead of whole sentences changed to one word. Zip codes were extracted from the address using RegEx. Subjects

were divided into categories 1-4, since there were four target variables in the multiregression. Subject was an ordinal, categorical variable. All other subjects were added to category 5.

In Appendix B, figure 2(a) shows Features on the x-axis and the Percentage of Missing Values on the Y-axis. The records that were missing in questions were deleted. If the zip code, sun exposure, and type has missing values, the values were replaced by the mode, as they are categorical variables. To eliminate the typos, more stop words were added other than punctuations and normal stop words (Appendix B, figure 2(b)) in English. Followed by the creation of word could (Appendix B, figure 2(c)) from the questions attribute with the most frequent words from the questions asked.

The categorical variables were encoded using three techniques such as cyclic, one-hot, and target encoding as shown in Table 1. Different variables were encoded using different techniques. For text analysis using Python, two types of multi-logistic regression models namely, one vs Rest (OVR) and multinomial were used. In Appendix B, figure 2(d) shows the confusion matrix obtained from both the models. The confusion matrix was the same for both the models. Using a confusion matrix, the precision, F1-score, and support was determined to examine the accuracy of the model as shown in table 2. The accuracy of both the models were similar and was approximately equal to 47.0%.

Table 1 showing different types of encoding and attributes categorized according to the encoding

Cyclic encoding	One-hot encoding	Target encoding
Month (12 unique values)	Type (2 unique values)	Zip code (Multiple unique values)
	Sun Exposure (3 unique values)	

Multi-logistic model helped in predicting the classes (Trees or Shrubs, Flowers, Fruits or Vegetables and Diseases or Insects) by month over the zip-codes of Mecklenburg County. The program manager will get answers for which questions they should prepare volunteers for.

However, “what type of questions related to each subject can be asked”, will be answered by topic modeling. Questions can be further categorized in topics such as maintenance, diagnostic, or conditions. The most frequent topics, such as diseases, insects, flowers, trees, shrubs, fruits, and vegetables obtained from topic modeling in Python (Appendix C, figure 3(a)) were associated with Zip Code in Mecklenburg County.

Table 2 showing outputs of all classes and One vs Rest (OVR)

	Precision	Recall	F1-score	support
1	0.38	0.67	0.48	142
2	0.20	0.01	0.02	84
3	0.23	0.04	0.07	68
4	0.00	0.00	0.00	60

The data set has information on the amount of sun exposure per day. Figure 3(b) in Appendix C shows the relation between the frequent topics and the sun exposure with zip code. This allows the program manager to check the most frequent issues in an area to reduce the number of travel hours. With this information, the program manager is able to customize the monthly newsletter according to the problems faced in an area.

Volunteers and Donations

Using Volunteer data, our goal was to improve work efficiency of the volunteers and help to retain and increase the number of volunteers. Appendix D, figure 4(a) shows the top 10 volunteers in the Master Gardener Program. There are a few volunteers who are regular to the program. However, most of the volunteers are not significantly contributing towards the growth despite free intensive training provided by the Master Gardener Program. There is an outlier in the number of volunteer hours (21.5 hours in a day). Appendix D figure 4(b) shows a declining trend in the number of volunteer hours from the first quarter to the last quarter of the year. Quarter 1 is the growth season where the Master Gardener Program receives maximum queries to be solved

and the training of the volunteers is done. Appendix D figure 4 (c, a) shows the inverse relationship between number of travel hours and number of volunteer hours which reduces the efficiency of the master gardener program. It can be seen that there is a repeated trend (every 12 week or 1 quarter) in which volunteers are contributing their time to the program. From figure 4 (c, b) in Appendix D, it is interesting to note that volunteers are working more during the weekdays rather than the weekend which poses a question of correctly logging the number of hours by volunteers. Some of the volunteers are working in more than two programs within the Master Gardener Program (Appendix D figure 4(c, c)).

Although the Master Gardener Program has not received a lot of donations from 2019 – 2022, by analyzing figure 4 (d) and figure 4 (c, a) found in Appendix D, it can be interpreted that customers tend to make donations after 3-4 weeks of service. This is a valid reason as an individual will wait to see results because rewarding the service and in case of plantation, 3-4 weeks is a fair amount of time to observe the reasonable results.

Recommendations & Conclusions

Recommendations based on text analysis:

Editing the Google form to require all fields of information. The current form makes several choices optional, which means that the organization is not getting access to all the data that they need. Additionally, the program should also require registration on the program's website for those using their services. Both of these will provide the organization with more data on the people using their services, as well as the problems they are facing.

Using the rankings obtained from the text analysis that we ran, creating targeted workshops, online classes, newsletters, and similar educational content. It can cover an entire category, such as trees, or focus on a specific topic. Similarly, by putting together relevant gardening terms the program can prioritize its services at the topic level. For example, in our

analysis, birds were found to be the least common issue, so content related to birds can stay minimal in educational material.

Recommendations based on volunteer and donation analysis:

We recommend assigning the responsibility of events to volunteers that live closest to the area. This will cut down on the number of travel hours, which is inversely related to volunteer hours. As noted above, some volunteers only help in one program, while other volunteers help in multiple. Distributing the volunteers into the different programs as needed will improve the benefit that the Master Gardener Program is getting out of the volunteers. Currently, training is conducted during January, February, and March. This is the time that the program sees the largest number of volunteer hours and the largest number of questions from the public. Additionally, many crops need to be planted in March or April in North Carolina. We recommend that the program moves the training to October-December, so the volunteers are already trained during their busiest part of the year.

There is a trend in volunteer hours that shows many stop volunteering after a short period of time. To combat this, we recommend setting-up a system that would reward the volunteers for their time. It would make the volunteers more engaged in the program and make them feel more valued, increasing the likelihood that they will continue with the program. The program would only need to do something small, such as volunteer of the month or first choice in assignment, but it would be a way to thank the volunteers and let them know that they are appreciated. The program would also benefit from partnering with a hardware store, like Lowe's and Home Depot. It could potentially provide discounts, donations or advertisements. One specific example is the Lowe's Foundation. The Lowe's Foundation is a charity that has operated in Charlotte since 1957. It

focuses on smaller, local businesses or nonprofits and donates resources, with the intention of bettering Charlotte.

Recommendations for further analysis:

We recommend that the program manager investigate how hours are logged and pay closer attention to the hours that volunteers are recording. Some of the hours that volunteers are logging does not make sense and further investigation into the matter would help understand what the problem is.

The data shows that volunteers are no longer volunteering or volunteering less as time goes on. There is no known reason for this, but it does seem to be a trend among several volunteers. Since the program does spend so much time training the volunteers, it is a drain on resources for the program to be continually losing trained volunteers. Further investigation and analysis would explain the behavior and potentially provide a solution

Acknowledgement

We would like to thank Steven Capobianco, County Extension Horticulture Agent for providing us the data for this project and providing constant support on the type of insight that will be useful for the growth of the program. We would like to thank Prof. Booshra Ahmed and TA Loren for successful completion of the project. Lastly, we would like to acknowledge the team effort of each team member for support and significant contribution.

Bibliography

100 Gardens Club. (2022). *About the 100 Gardens*. About 100 Gardens - 100 Gardens. Retrieved May 4, 2022, from https://100gardens.org/about-100-gardens/

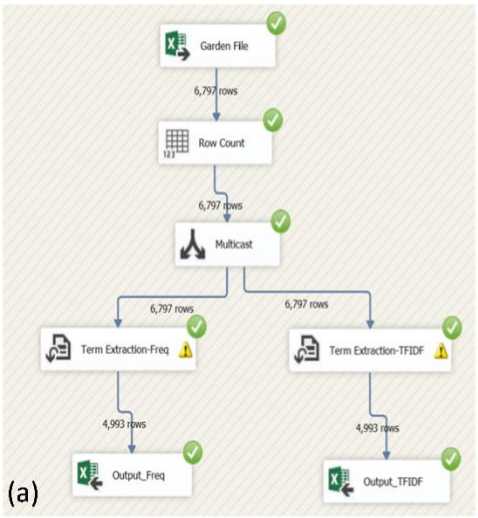
American Horticultural Society. (2022). *AHS Overview*. American Horticultural Society. Retrieved May 4, 2022, from https://ahsgardening.org/about-us/

Charlotte Council of Garden Clubs. (2020, May 11). Charlotte Council of Garden Clubs. Retrieved May 4, 2022, from https://charlottecouncilofgardenclubs.blogspot.com/

NC Cooperative Extension. (2022). *Extension master gardener℠ volunteers*. NC State Extension News. Retrieved May 4, 2022, from https://www.ces.ncsu.edu/categories/lawn-garden/extension-master-gardener-volunteers/

NC Cooperative Extension. (2022). *NC State Extension*. NC State Extension News. Retrieved May 4, 2022, from https://www.ces.ncsu.edu/

Appendix A:



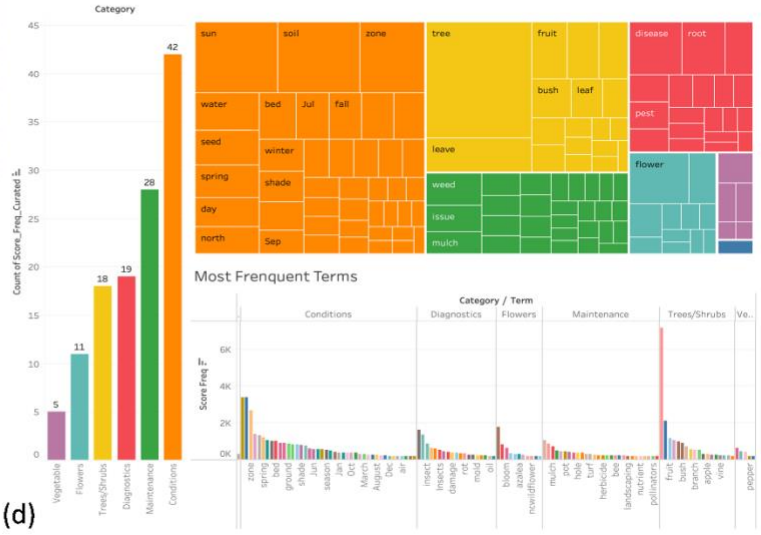
(a)

Category	Total Ranking	Relative %
Birds	312	0%
Vegetable	1850	2%
Flowers	5212	7%
Diagnostics	9494	13%
Maintenance	9627	13%
Trees/Shrubs	17754	24%
Conditions	31208	41%

(b)

Term	Score	Category
flower	1790	Flowers
rose	823	Flowers
bloom	653	Flowers
perennial	329	Flowers
camellia	294	Flowers
azalea	285	Flowers
bulb	251	Flowers
Flowers	208	Flowers
ncwildflower	200	Flowers
Cambria	196	Flowers
lily	183	Flowers

(c)



(d)

Figure 1 showing (a, b, c) results from SSIS package, and d) Tableau dashboard

Appendix B:

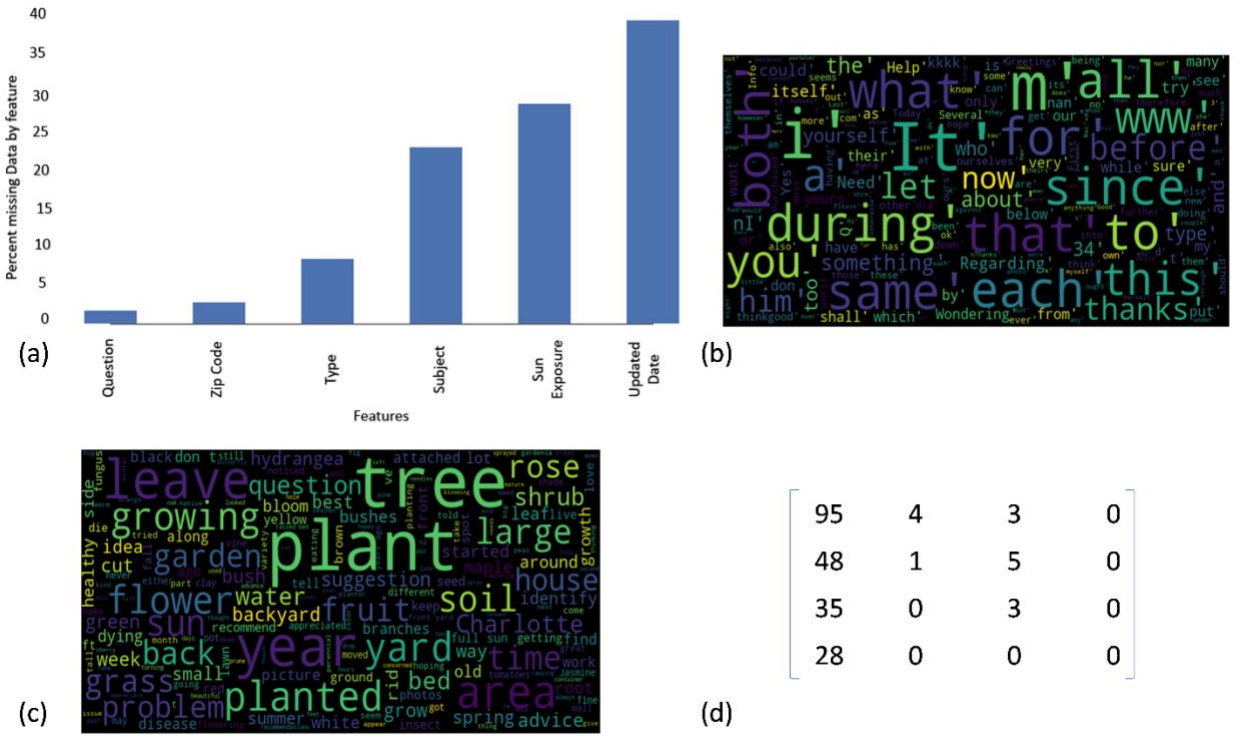


Figure 2 showing (a) percent missing data by feature, (b) stop words, (c) most common words, and (d) confusion matrix using one vs rest and multinomial method from text analysis.

Appendix C:

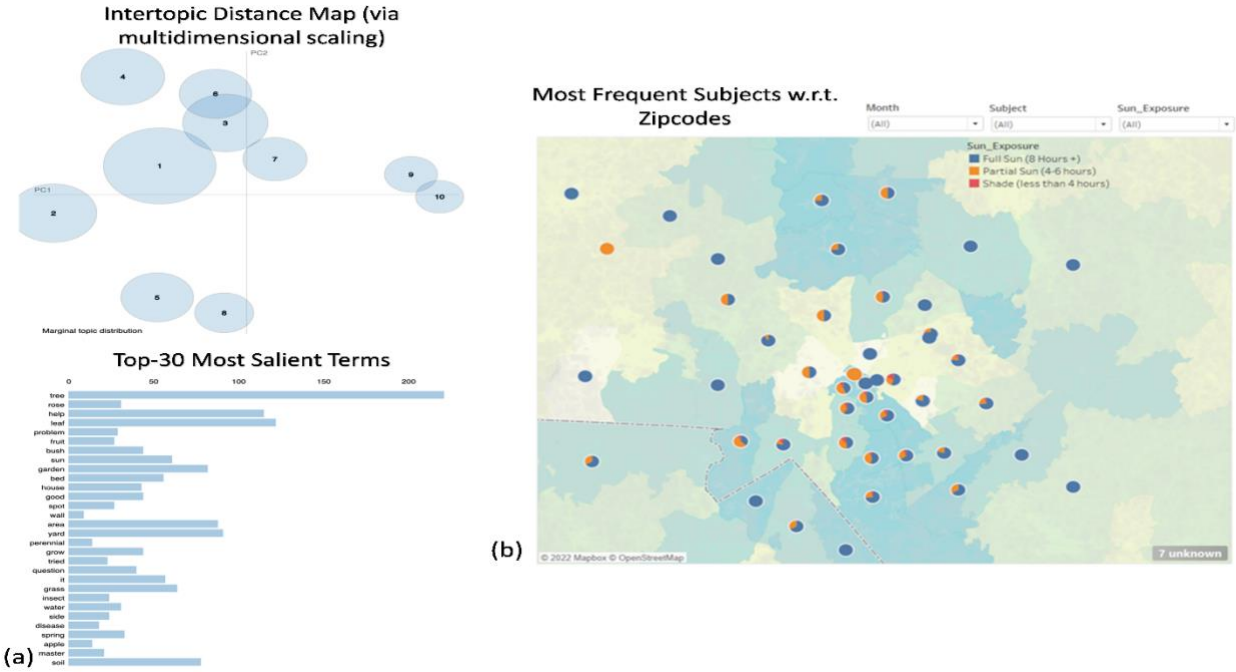


Figure 3 showing (a) overall topic model with ten topics obtained from topic modeling, and (b) a distribution of the most frequent topics obtained from topic modeling and the amount of sun exposure per day w.r.t. zip code in Mecklenburg County.

Appendix D: Volunteer and Donation

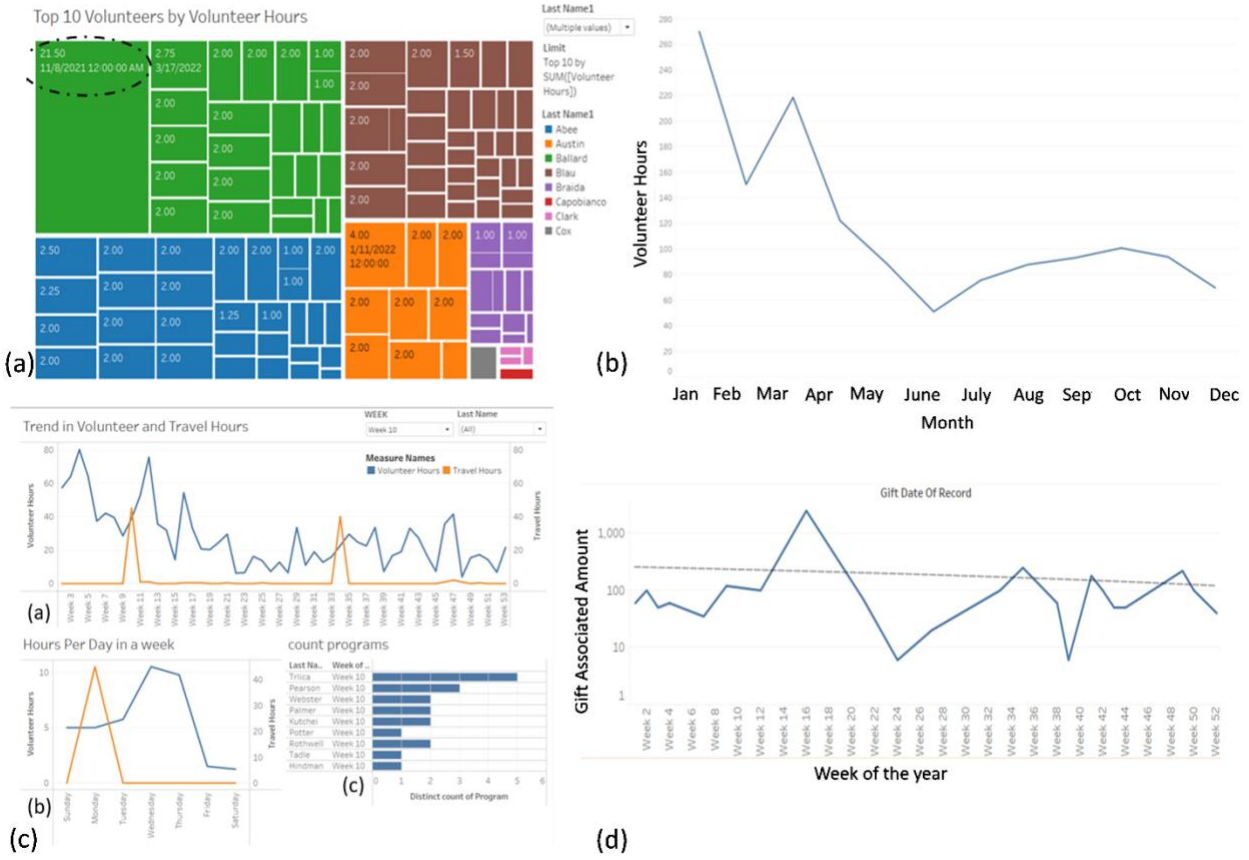


Figure 4 showing (a) top 10 volunteers in master gardener program by number of volunteer hours, (b) trend of volunteer hours w.r.t. to each month in the year, (c) snapshot of Tableau dashboard showing (c,a) number of total volunteer and travel hours that volunteers have recorded per week, (c,b) total volunteers and travel hours recorded on each day of week 10, and (c,c) count of programs in which each volunteer have worked in week 10, and (d) weekly trend in donation to the master gardener program

Appendix E:
Extension Center Resources

Extension has been a leader in making educational material available on the Internet. Many Extension publications and departmental leaflets are available online. Review the following websites to find horticulture and MGV information online:

- North Carolina Extension Gardener Handbook
- NC State gardening information
- NC State Extension Gardener portal
- NC State Extension Horticulture
- North Carolina EMG Intranet
- North Carolina EMGV portal
- N.C. Cooperative Extension Service
- NC State Department of Horticultural Science