# Deliverables, Goals, and Etcetera

This document will cover exactly what I hope to accomplish in my time researching on this project. This project is one I have been working on for the past several months, and I hope to continue throughout my time in graduate school. I’ve been working on a validation study on Kiri Wagstaff’s paper on constrained k-means clustering algorithms. I will include a copy of this paper, and any relevant works, along with this document on time of submission.

In my initial testing and original presentation on this project, I’ve been working on just trying to figure out exactly what constrained k-means clustering means, and how it is actually useful. I ended up getting perfect marks on my publication and presentation on this from the SWOSU Research Fair last semester. I will be presenting my next publication at Oklahoma Research Day.

I will also be including a link to my GitHub repository that will be linked to SWOSU when I am done.

This repository will include:

* The scripts that I’ve used
* A guide to create what I’ve created
* Any necessary papers that I’ve used
* Any other prevalent information that I may think necessary

The reason for this is twofold. One, it will hopefully bring SWOSU more recognition as a researching school. Two, it will be of help to anyone who will be studying this topic in the future, as there remains to be any proper guide on how to get started. I would be more than happy to submit my weekly reports to you as well as Dr. Evert and Mrs. Baugher. Please just let me know what I need to do in this regard. As for what my end goal will be, I’m not sure as I will be working on this for the better part of the next few years. I have talked to Dr. Evert and we reached out to Dr. Wagstaff and received her blessing to be doing this project. I think that my end goal for this semester will be to have two more publications under my belt, and I hope to have a working Python script to be able to fully complete this project. As of last week, I have two separate Python scripts. One that is working to complete an example of constrained kmeans, and one that is working on reading a DATA file into an array. I’ve been working with Mr. Smoot and Dr. Evert with trying to get this operational.

In the past week, I’ve gotten a successful figure to be printed in my Python script (Appendix A).

This was mainly due to me giving up on the Windows 10 operating system and moving everything into a Linux environment. I’ve also gotten a running script to be able to fully read the DATA file and put it into a multi-dimensional array. I will be working this following week on trying to be able to get the two scripts to work together. After that, it will just be a lot of testing, in theory. In addition to these, I have been working on creating a detailed guide on how to replicate the steps that I have taken.

I believe that this covers what I’m working on, what I want to do, and what I’ve worked on in regard to this project. Now I will go into exactly why I want to do this. I feel as though my time with NASA has been immensely fruitful, leading me to be fully inspired to pursue a career with NASA. There are few places that have been as inspirational as NASA, and I hope to, quite literally, shoot for the stars. I’ve been working in the NASA L’SPACE Academy in addition to this, and it only furthers my drive to be employed with them in the future. Dr. Wagstaff also works for NASA, stationed at the Jet Propulsion Laboratories. I hope by doing a validation study under someone so close to where I want to be, it will give me a direct line into the system once I am done.

Also, I’ve always been fascinated by artificial intelligence. This project will tie nicely into my research into the field of artificial intelligence by eventually creating something that will be able to do these problems with ease, on its own. I hope to make the science fiction of my youth something of a science fact for tomorrow’s youth. This project is just a stepping stone in order to get me there, and I’ve loved every bit of all of the problems I’ve ran into. It’s a puzzle that only will take time in order to fully crack.

Update as of 3/14/19

During the past several weeks, I’ve been working with both Dr. Jeremy Evert & Mr. Devin Smoot in order to create a proper Python script in order to be able to take in any CSV/Excel dataset in order to find the *k*-means of it.

In addition to this, I was present at the Oklahoma Research Day in order to present my findings. I was able to get my research poster published via SWOSU. I’ve included both of my posters in the works cited page later on in this document. It was an absolute blast to be able to present my findings at ORD, and I had several fantastic conversations with people about it. I was very proud at how my poster ended up turning out.

As of right now, I’ve been struggling with the specific way I’ve been trying to figure out exactly how to get the formatting right. I’ve been encountering constant errors with trying to import an excel file into a script. The biggest issue with the dataset that I’ve been using is the fact that it is missing several datapoints. These points are in the data file as a ‘?’ instead of any point, leaving it as null. This was throwing countless errors in my program.

I’ve also had issues with importing a .data file into an Excel-friendly format. That was pretty easy, after I figured out that the installation of Excel, I had didn’t support it naturally but after I did a fresh install, it was freely able to take in the .data file (it read it as a .csv file) and convert it to an excel file. If needed, I can include the excel file that I’ve used for my dataset. The dataset I’ve been using is the Soybean dataset, found at [https://archive.ics.uci.edu/ml/machine-learning-databases/soybean/.](https://archive.ics.uci.edu/ml/machine-learning-databases/soybean/)

By far the most annoying issue has been trying to get tkinter to work in the way I’ve wanted it. In Python, there’s a way to have tkinter throw an open file dialog, but I wasn’t able to get it to open and accept without crashing due to several different issues. What ended up happening was it was reading it in as a photo and not an excel file. As for why, I genuinely haven’t been able to figure it out yet. The way that I ended up figuring out a workaround for now is to put it into a try statement. This has the added annoyance of popping up for input twice, but it’s working, so I’m going on with it for now. I’ve been able to verify that it is reading all of the excel file into an array, so that’s pretty much what I wanted it to do to be able to run the sklearn library for kmeans.

Other than this, I don’t have much of anything to say on my progress. If there is anything else you need, please let me know. Thank you.

# Appendix A Figure

1:

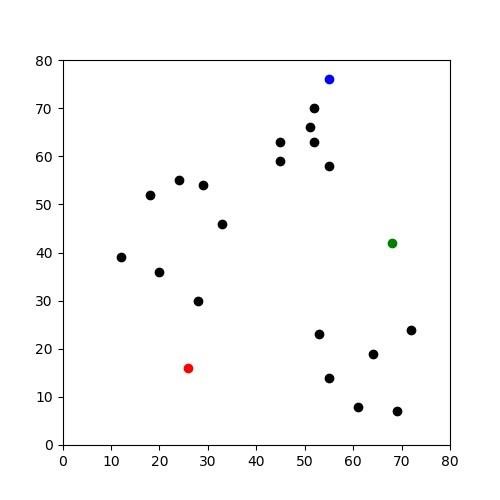


Figure 2:

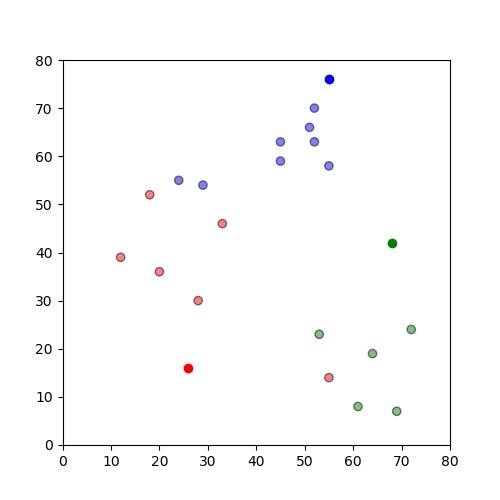


Figure 3:

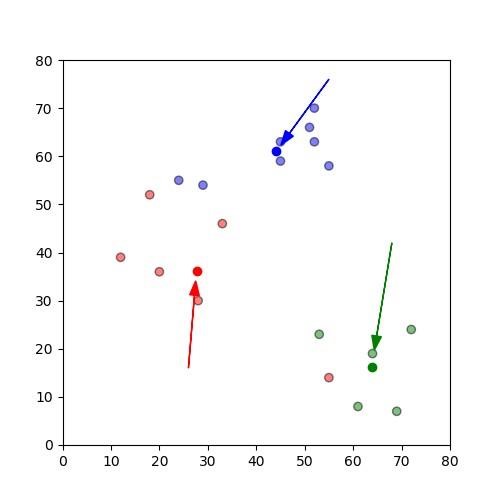


Figure 4:

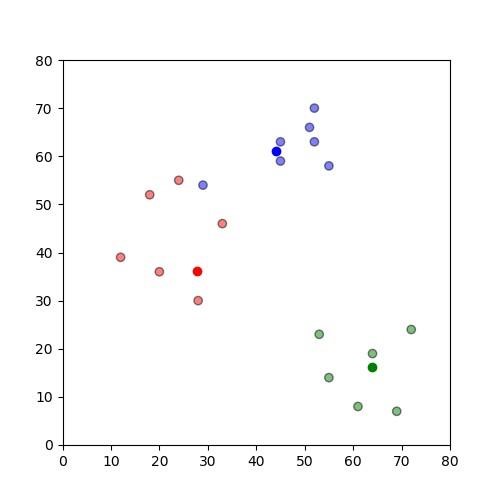


Figure 5:

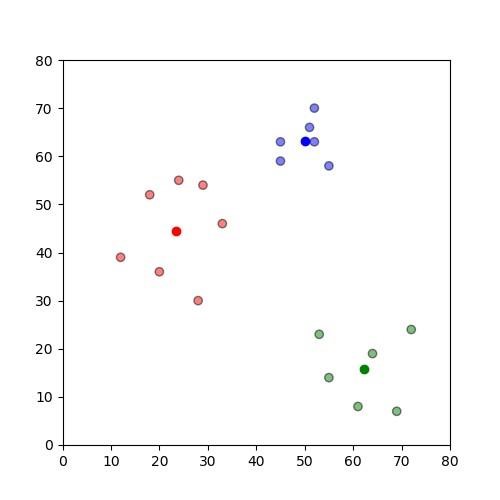
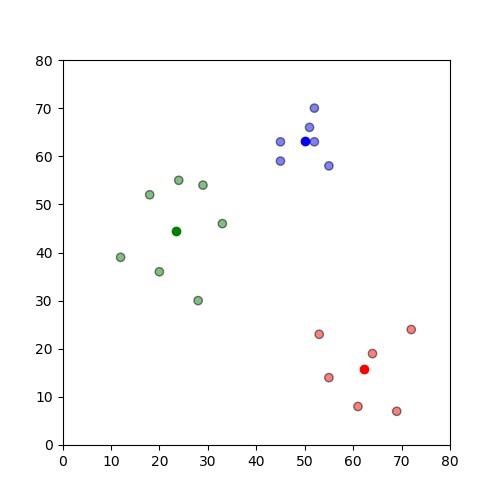


Figure 6:



Appendix B: Start of the Python File



Figure 1. The start of the GUI

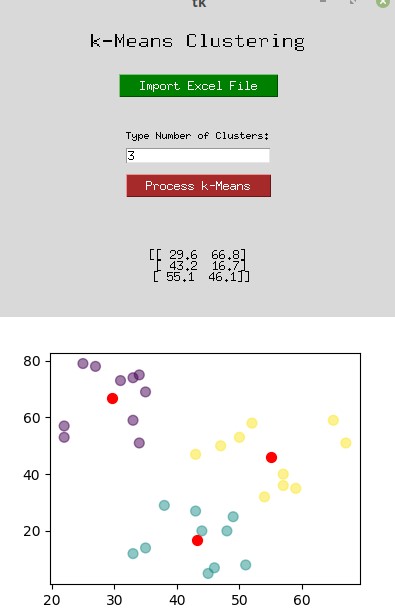


Figure 2. The first wave of clustering.

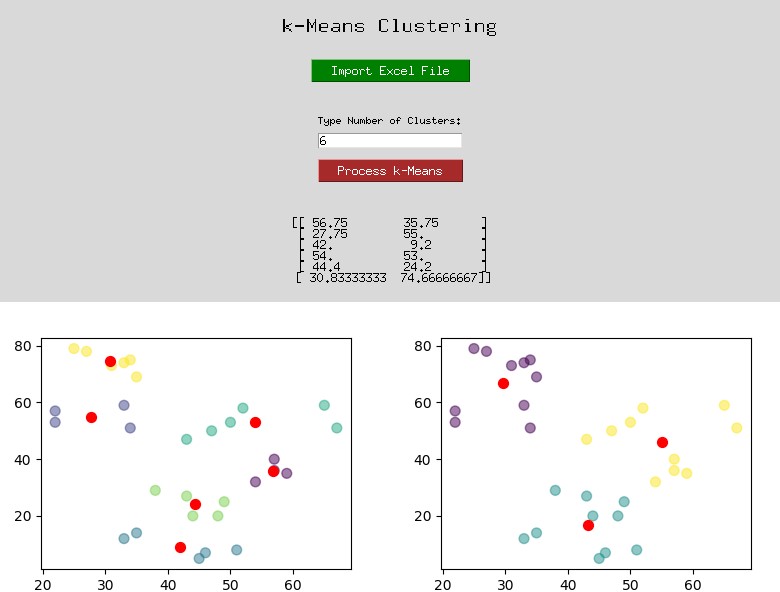


Figure 3. The example showing multiple different runs of clusters.

# Works Cited

Basu, Sugato, Mikhail Bilenko, and Raymond J. Mooney. "A probabilistic framework for semi-supervised clustering." In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 59-68. ACM, 2004.

Bradley, P. S., K. P. Bennett, and Ayhan Demiriz. "Constrained k-means clustering." Microsoft Research, Redmond (2000): 1-8.

McDaniel, Nicholas; Burgess, Stephen; and Evert, Jeremy, "Constrained k-Means Clustering Validation Study" (2019). *Student Research*. 20.

<https://dc.swosu.edu/cpgs_edsbt_bcs_student/20>

McDaniel, Nicholas; Burgess, Stephen; and Evert, Jeremy, "Constrained K-Means Clustering Validation Study" (2018). *Student Research*. 12.

https://dc.swosu.edu/cpgs\_edsbt\_bcs\_student/12

Wagstaff, Kiri, Claire Cardie, Seth Rogers, and Stefan Schrödl. "Constrained k-means clustering with background knowledge." In ICML, vol. 1, pp. 577-584. 2001.

Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." IEEE Transactions on neural networks 16, no. 3 (2005): 645-678.