

Write up

Project planning

My original inspiration for this project was the fact that my entire family played a role in treating patients during the COVID-19 pandemic. Each of them worked in different sectors, my father as a doctor, my mother as a nurse, and my brother as an EMT. The data set I chose is from data.gov, and it shows the conditions and contributing causes listed along with deaths from Covid by age group and where the death occurred. I had to create the edges, vertices, and graph all from this data by taking multiple columns into account to ensure I was getting unique cases. I chose to apply clustering logic to this data set, because I assumed that it would be interesting to see how each instance of Covid deaths from this data compares to what has been produced by doctors and epidemiologists since the pandemic.

cluster.rs file

In my cluster.rs file, I create a Graph struct that allows me to construct a graph using the edges collected in my main.rs file. I built a function to create an adjacency list that is used to create the graph. I ran into an issue because the labels that were created from the original data set weren't numerical so I had to convert the labels to numbers using a HashMap. From there, I create a function that is essentially a k-means clustering algorithm, that creates clusters using the adjacency list function and creates centroids and assigns vertices to clusters, and then calculates new centroids and continues from there. Finally, I implement a function called 'find_best_k_silhouette' that implements the silhouette coefficient as a means for finding the best k for my data. Since I have been taking DS210, and DS121 at the same time, I recalled from earlier this semester where this was learned, and decided my best bet was to translate it from how it was implemented in Python but into Rust.

main.rs

main.rs includes both cluster.rs and test.rs as mods and sets the stage for the entire project by reading through my imported file, parsing through each line, and creating a list of edges that are then used to create the graph. I ran into a bit of an issue here because at the last second I realized that my edges were not compiling in a way that expanded the amount of vertices needed for the project, which is ultimately a failure in my project, but because I didn't realize until it was too late and I didn't want to risk having to start from scratch, considering the various delays I have had in general.

test.rs

In this file I run tests on the adjacency list function, which I had to follow the same pattern of using three Strings in a vector because that was what I implemented when I was originally constructing edges. In hindsight I realized that I would've been better off choosing a better data set that would've been easier for me to create edges from so I didn't have to adjust so much in this project. I also created a test for the find_best_k_silhouette which required me to create a fake graph, and set a silhouette coefficient that is used in the test.

Results

I don't have much to show for my results, because I ran into multiple errors within my code. I'm not really sure where I went wrong, but in all honesty, I can say I was perhaps too ambitious with this project. I had to search for additional resources to help for addressing issues that I kept running into (for example when I was making the find best k function, I had to use find a way to respond to the way I crafted the edges originally and I ended up having to import the euclid library).