Nicole Aguilera
DSC 680 - Milestone 2 Check-in

1. I found it surprisingly difficult to access the data in the way that I was originally going to. I was planning on obtaining my personal data on Spotify because they have an API that the public is able to access when an account is linked. However, I was not able to access the API site through the keys that I was given when registering. I ended up with an error stating that one of the keys was not valid even though I obtained it from the site itself. Nonetheless, I was surprised at how willing this company was to work with developers through this API because not all companies are as willing.

2. The data set was not what I thought it was due to the troubles I was having with accessing the data through the API. This caused me to switch mind sets before beginning the project to acquire some sort of data. Searching on Kaggle, I was able to find a data set of random songs from 2017 that had the elements that I was looking for, which were the attributes of the songs. In this way, I was able to still continue with the original plan of model that I had for the project, but just using different songs from Spotify.

3. I have not adjusted my research questions or approaches to the project in question. Even though I had to change the way in which I found the data for songs from Spotify, this did not alter the approach I was taking for this project. I am still running a clustering algorithm of k means to find how many clusters the algorithm puts those songs into, and the correlation between the metrics of the songs in the data set. The correlation matrix will be constructed using a heatmap, through which all of the variables, in this case, the metrics of the songs, will be displayed in how they correlate with one another.

4. My method for this clustering problem is working fairly well. I have always thought that k means is a pretty straight forward approach to clustering data, even though we as humans do not always understand how the algorithm is grouping the data together. The model itself is not extensive and the data just needs to be clean in order to apply the model to it. I also used the elbow method in order to choose the number of clusters, or k, to be used in the project. This method plots the variation in the data and looks at where there is an elbow in the curve. That point determines the number of clusters to be used for k means.

5. The challenge with this project was continuing on with the API through Spotify itself. As I had mentioned before, it was difficult to get my code to gain access to the site. So, after hours of trying and looking up solutions across the web, I decided that it was in my and the project's best interest to shift gears to a predetermined data set. I know that this project would have been more insightful to me, personally, if I were to use my own data. However, with the time restraints that I have for this project, it was better to see how the model preformed rather than the exact songs that were in the data set. I can

always go back in the future and run the same model with my own songs once I figure out how to gain access to Spotify's API.