Introduction

For this practical session, we use clustering to

Implementation

We decided to just implement clustering in Python without KNIME. Immediately we ran into issues with the size of the the initial database. When looking at the csv our PyCharm IDE, the whole IDE would freeze. After realizing that this was our problem and not Pythons processing ability, we moved on. Our first step was to build date taken and date upload from the current columns in the dataset. Since it was not well represented initial, this was to be done first. After our new columns replaced the old ones we still had too many columns. As pointed out by someone in our class, there were a lot of repeated columns. This was easy to add in and made our dataset go from 167702 data points to just 13504.

Now we can implement clustering. We started out using the K-Means clustering algorithm. Once we ran the code, we waited and waited and waited. Clearing there was too many iterations and too much data to be processed. We tried minimizing the data as much as possible however, nothing was fixing the computation time. Luckily, there was MiniBatchKMeans which splits the dataset into small batches for faster computation. Based on the sckit API was was little to no different between the two implementations. After we were able to let the clustering algorithm run efficiently, we then plotted our data. We started off our clustering by using the latitude, longitute, and upload date and time. However, this gave us some weird results on our graph. The clusters were very undefined and it seemed like it wasn't working. Then it was clear that including the upload and taken dates were causing the undefined clusters. We removed them from the computation of the clustering, and we just used latitude and longitute. This resulted in much clearer results. We could see clear separations of our data. But at this point, we were plotting the values to a blank plot. But these were coordinates on a map, so we took a picture of a map that was defined by our coordinates and used it as the background to our scatter plot. This was not the best solution but it was good enough for us to interpret the data as a location. Once we had completed our program, we then created multiple graphs using different sized clusters for analysis.

Results

Explain our results!