Khawar Murad Ahmed, Shreeman Gautam
April 6, 2022

Peer Review

Group 2

The main contribution of this project is the evaluation of clustering algorithms based on the given data. Evaluation, specifically, rests on drawbacks, benefits and accuracy of each algorithm. It would be nice if the terms drawback, benefit and accuracy could be defined loosely for each algorithm. The project idea is novel because categorical values are converted to numerical values(using probability distributions, wassterstein distance) and as is known, clustering is usually done with numerical values. At first glance, especially with 31.8 million records, the plan might seem unreasonable, but as mentioned, the algorithms were tested on all the records and further, GPUs will be involved in the future to expedite the process. The idea of the project is to evaluate clustering and so far, the report has shown how the different clustering algorithms work in regards to the data without specifying what the drawbacks, benefits and accuracy of each algorithm are currently. We hope that the final report will make the evaluation clear.

Group 22

Currently, this project lacks a clear direction in terms of what information it is trying to extract from the analysis. Essentially As mentioned in class, evaluation is extremely important in terms of benchmarking how well the experiment worked. On that end, there are two components that this project needs. The first one is a clear, well-defined hypothesis. I think some glances at the data would help develop some intuitions regarding what the data looks like, and some expectations regarding how the algorithm will perform. Naturally, once that hypothesis is formed, coming up with evaluation metrics would be easy. It is extremely hard to evaluate how reasonable the project is and whether current results support the idea because the project seems to lack a main idea. The current results also don’t provide much information about the data in its current state.