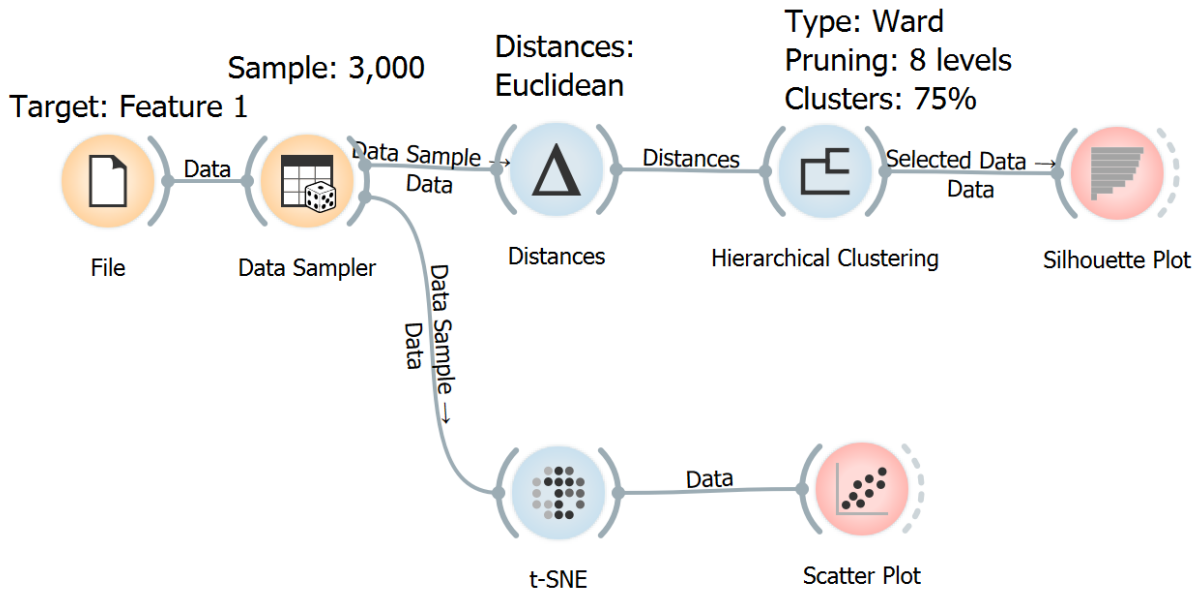
Jennifer Bochenek, Toby Obunadike, Joan Kibaara
Spring 2020

Introduction

For this project, we used the program Orange to review the MNIST handwritten 10 digit dataset and then the Movie dataset. For both of these, for ease of computing we reduced both to a sample of 3,000. Below, see the figure 1 for the workflow for the MNIST dataset.

Figure 1. Orange workflow for MNIST dataset.



Dataset- MNIST

After inputting the file, we made sure that every variable was numeric except Feature 1, which was the target variable. Using a sample of 3,000, we addressed two methods of clustering by running t-SNE and Hierarchical Clustering. In order to run Hierarchical Clustering, we first used the distances module which was set to Euclidean distances. For the Hierarchical Clustering, we used the Ward algorithm, and set pruning to 8 levels for ease of viewing. We allowed the program to select the clusters (4 total) by setting the height ratio to 75%. To check the quality of the clusters we also requested a silhouette plot.

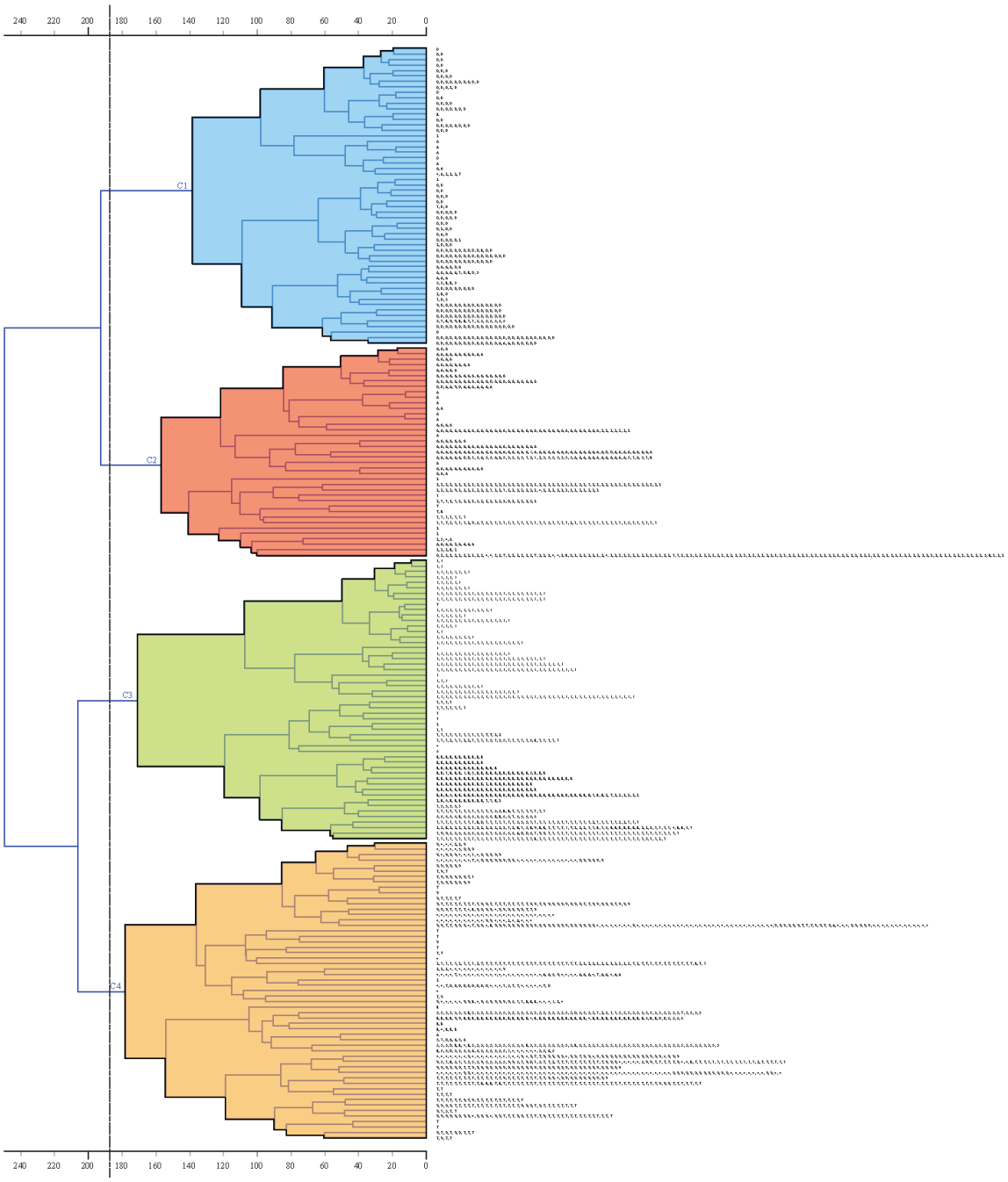
Hierarchical Clustering - MNIST

Figure 2 (also attached as a pdf) below is the results of the Hierarchical Clustering with 4 clusters. The four clusters ended up being C1 which was predominantly 0s (blue) with some 6s and lesser amounts of other numbers, C2 which was predominantly 6s (red) with some 2s and then some small amount of other numbers, C3 which was predominantly 1s (green) which trails into 8s and then 3s, then the last cluster C4 (orange) which was an assortment of 'stick' numbers such as 9, 7, and 4. This last cluster is the least differentiated, which shows the difficulty in correctly categorizing the numbers based on this method.

According to the silhouette plot (figure 3, attached only), clusters 1 (blue) and 3 (green) were the best performing, while cluster 2 (red) was the worst performing. Cluster 4 had a number of items that

were positive (i.e., like the cluster it was assigned) but it also had a long tail that trended negative (i.e., was more alike another cluster). This was somewhat expected as that was the cluster of what we identified as ‘stick’ numbers, but it also had an assortment of ‘round’ numbers in it as well.

Figure 2. Hierarchical Clusters of the MNIST Dataset.

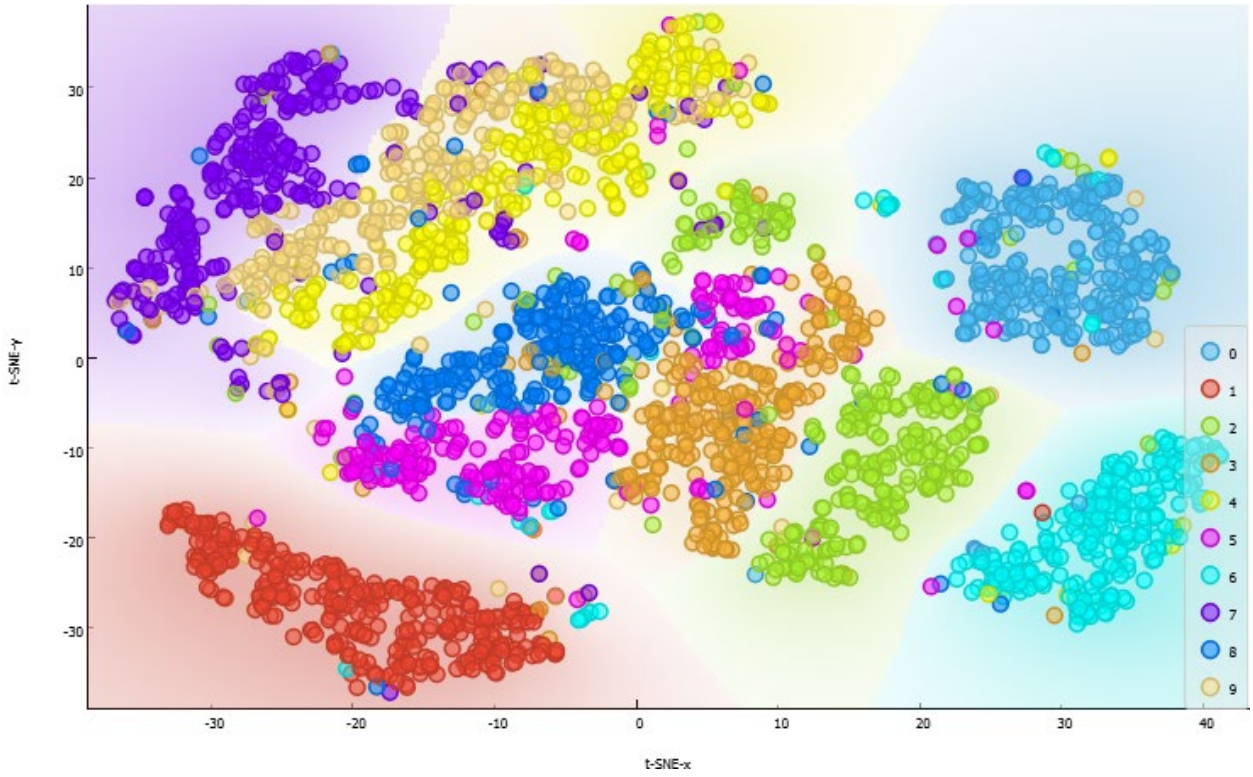


t-SNE - MNIST

We also ran t-SNE, also called t-distributed stochastic neighbor embedding, which is appropriate for high dimensional datasets such as the MNIST dataset. Figure 4 shows the clusters assigned by this

method. In the graph we can see that there are several clusters, but three stand out separate from the rest. These are the blue cluster in the upper right, the aqua cluster in the lower right, and the red cluster in the lower left. The remaining clusters were not as distinct from each other, but the purple cluster in the upper left, alongside the tan and yellow cluster are slightly set apart from the middle cluster. The clusters mostly align with a particular number, with the blue as 0, red as 1, aqua as 6, purple as 7 and tan and yellow as 9 and 4 respectively. The middle cluster is made up of the rest of the numbers: 2, 3, 5, and 8.

Figure 4. t-SNE of MNIST Dataset by Target Variable



Interesting Clusters - MNIST

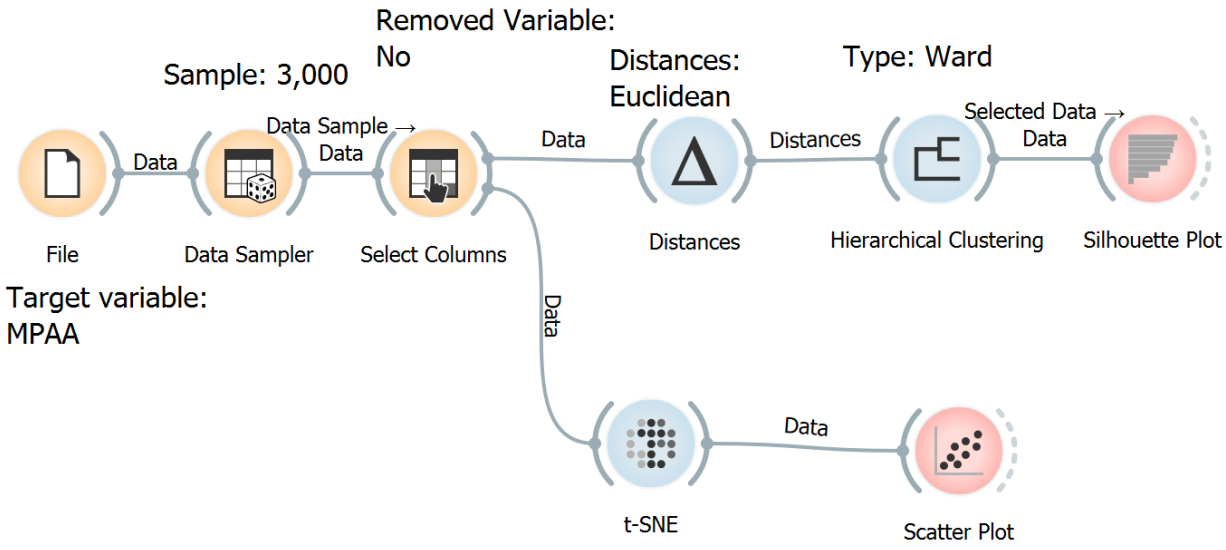
Two of the interesting clusters, to us, were the way that 9 and 4 merged with each other (tan and yellow in the t-SNE graph), and how unique the number 1 was in both methods. The merger of 9 and 4 in the t-SNE results was interesting because we saw the same results in C4 in the Hierarchical Clustering. Here we also see that, like in C4, the number 7 (purple) is close as well, while 8 (dark blue) is also close. It is just interesting to see similar results for both methods, despite using slightly different approaches. Like the Hierarchical Clustering, the 0s (C1, or light blue), 6s (C2 or aqua), and 1s (C3 or red) stand alone with only a few intrusions. 1 (C3 or red) was particularly interesting because it was one of the better performing clusters based on silhouette plots, which is reflected in the t-SNE graph by the distance it has from the other clusters.

Dataset- Movies

We used a similar method to review the movies dataset, as seen in figure 6 below. We set MPAA (rating for appropriate audiences) as the target variable because we were interested if there were clusters based around rating at first. We set the sample to 3,000 for ease of computing, and removed the variable

named ‘No’. The variable ‘No’ was a numeric ID variable that had no particular meaning and should thus be removed before calculating the clusters. The method for the Hierarchical Clustering and t-SNE were as before.

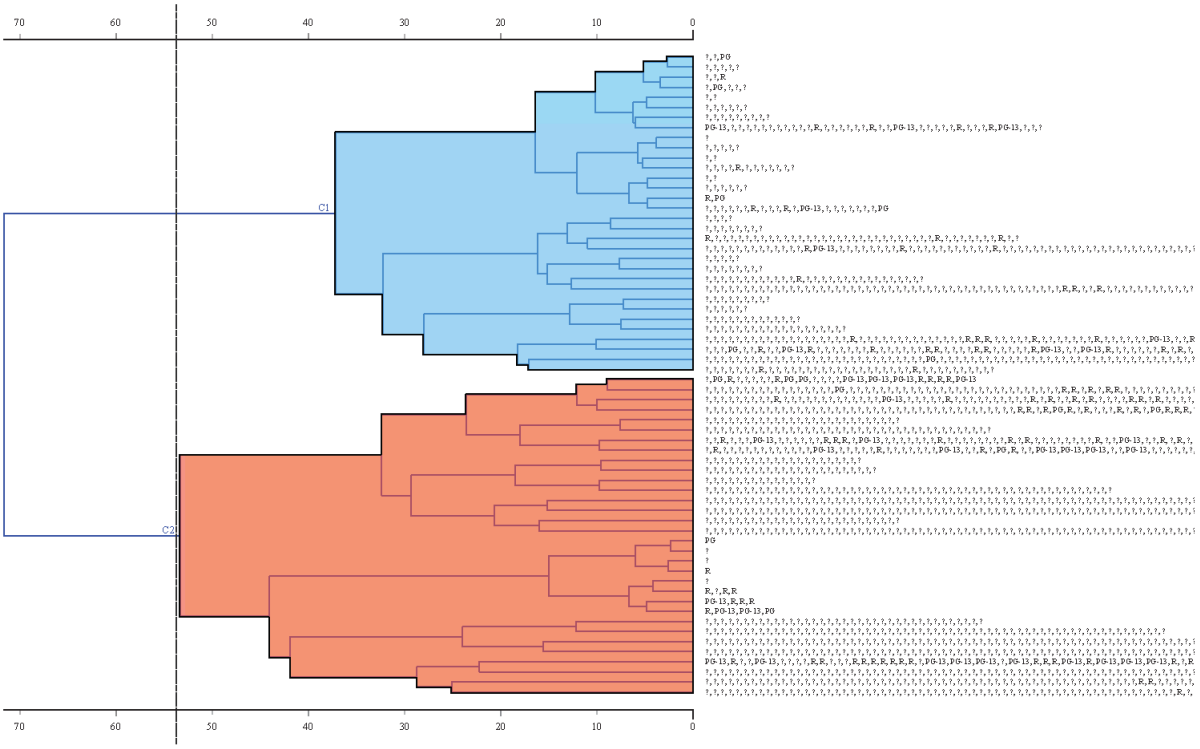
Figure 6. Workflow for Movies Dataset



Hierarchical Clustering - Movies

For the movies dataset, using the same criteria as for the MNIST data (75% of the data), we found two clusters. Both of them had unusual silhouette plots (figure 8, attached only), with C1 not having a high positives (near 0) but also only having a small number of negatively associated items in the cluster, C2 was the opposite, with a spike in positives at the beginning, and then have a long trailing negative that near 0. Ostensibly, that means that cluster 1 was not a good fit for its data, and that cluster 2 fit most of its data. What might perhaps better explain these results, would include using different cluster selection parameters to determine if there were additional clusters or more accurate clusters.

Figure 7. Hierarchical Clusters for the Movie Dataset



t-SNE - Movies

What those clusters became became more obvious once we ran the t-SNE. The graph shows approximately 9 clusters; however, we have reason to believe that the results are being influenced by the 7 dichotomous variables that sorted movies based on genre (e.g., action, animation, comedy, drama, documentary, romance, and short film). With these being dichotomous (i.e., the value is either 0 or 1) they would have a strong influence on the final distances between the groups, more so than the target variable of rating. Because each genre is each own variable, we present figures 9 and 10 to show the clusters “Short Films“ and “Romance “. Short films predominantly occupy the right portion of the map (shown in red in figure 9). It also bleeds somewhat into documentaries (the small cluster below), and is found in small amounts in other clusters. Figure 10 shows romance movies, which is located in the middle of the bottom of the graph in red.

Figure 9. Short Film Movies

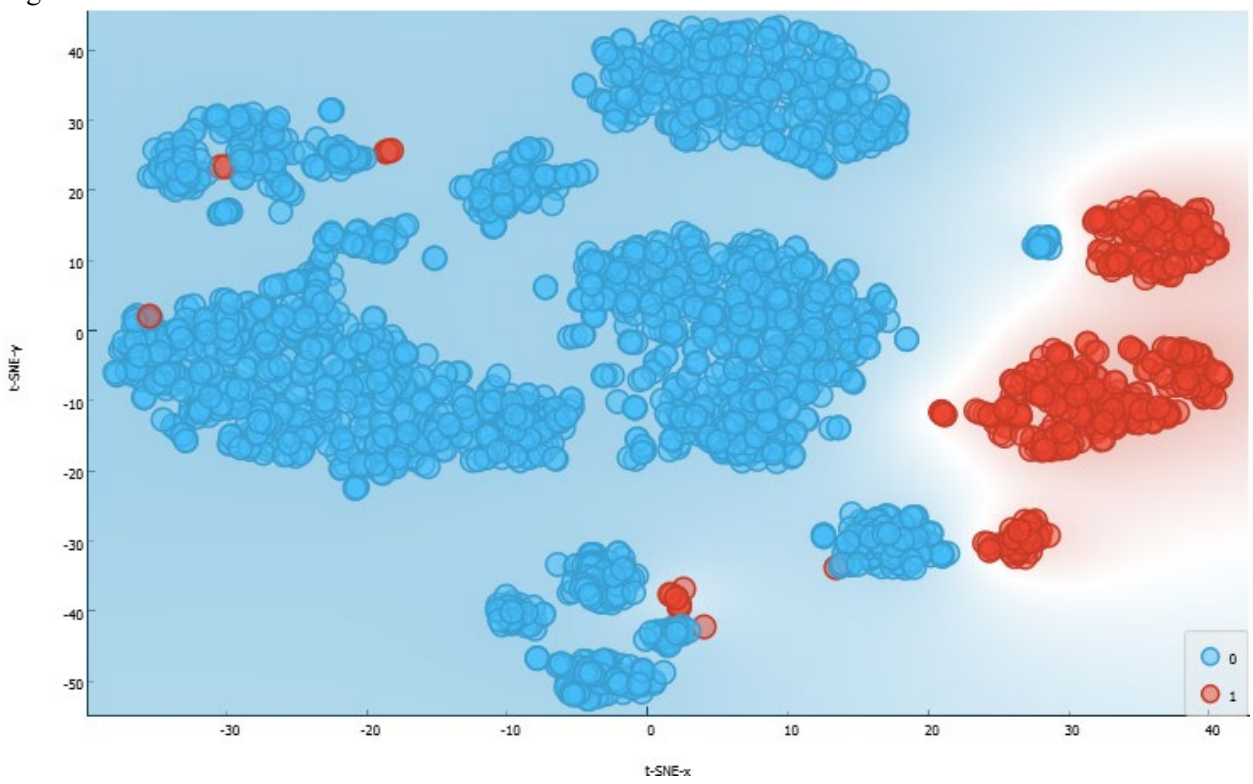
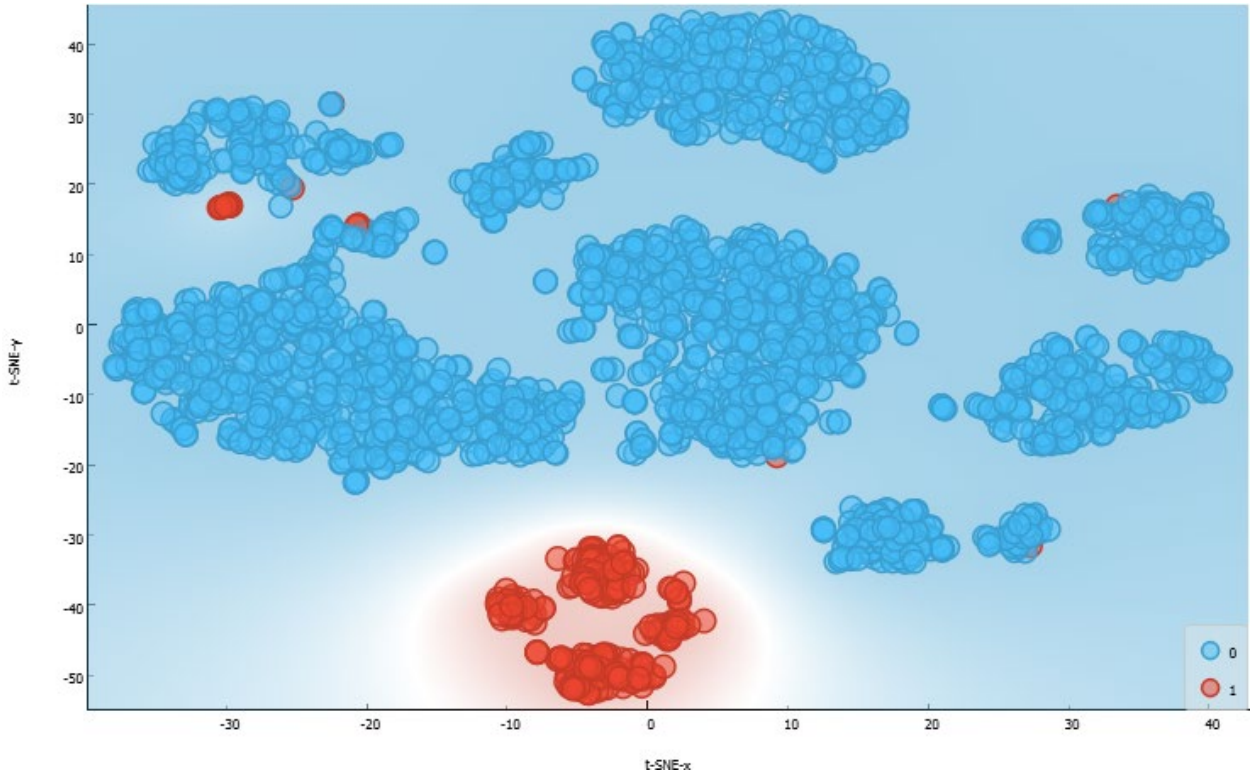


Figure 10. Romance Movies



Significant Findings - Movies

Since the categorical variables available in the dataset only allows us to view one genre at time, we chose to show short films and romance clusters. However, in each graph we can see that there are other clusters, they just are not named. Two of the most significant findings are that the Hierarchical Clustering was not as effective as the t-SNE, in that the t-SNE was able to easily show clusters and the Hierarchical Clustering did not. In the previous dataset, mostly made of continuous numeric variables, they performed equally well. However, this dataset was heavily influenced by dichotomous variables which t-SNE was better able to handle then Hierarchical Clustering. The second important finding was that the clusters using this dataset were particularly strong, in that they aligned very closely to the genre with only a few outliers. This can be attributed to the weight that t-SNE gave the dichotomous variables, but in comparison to the MNIST dataset results, there is less scatter among the movies dataset with regards to genre.

Our Experience

As a novice user of the Orange program, it took us a couple of days to successfully import the data into the system and set up the workflow. We utilized different resources from online and lecture notes from the class, when creating our workflow. We also did additional research on the t-SNE algorithm and silhouette plots that were used in our workflow. It was useful to have to troubleshoot and figure out how some of the models work, because we gained a greater understanding not only of the datasets, but the Orange environment and model requirements.

One of the project tasks was to use the same workflow for a different dataset. While we were able to successfully change our dataset and run the algorithms, we could not draw much useful information from the new dataset. We agreed that each workflow is unique to the data set it was created for.