Assignment-3 Visual Analytics 690V

First Name: Sanjay Reddy; Last Name: Satti; Spire ID: 31315112

Github Username: Sanjay-Reddy-S
Github Link: https://github.com/Sanjay-Reddy-S/Visual-Analytics-Assignments/tree/master/HW3

Tested on the local system in Python3 (All libraries should already be present in Anaconda)
The rendered interactive .html files are also present in the repository.

I've used K-Means and DBScan for clustering.

Besides online sites, I've also referenced Pearson's Introduction to Data Mining textbook

There are 2 files

- Two_Feature_clustering.py
  - Used only Region and Milk features to understand the algorithms in general
  - Although I've seen Silhoutte method, I understood the elbow-method better and hence used that to pick the number of clusters
  - For DBScan, directly using the values, resulted in all records being labelled as outliers. This particular aspect has also been described in the textbook, stating that "defining density for High-dimensional data is more difficult". So whenever DBScan is used, I've normalized the values by dividing it with the range of that feature
  - Once the variables are normalized, DBScan actually varied very less when changing its parameters like eps and min_points
  - Referring back to the same textbook, "K-Means has trouble clustering data that contains outliers and outlier detection can significantly improve in such situations" Using DBScan definition of density, I've actually removed the outliers and rerun the K-Means algorithm. I felt it returned more natural clusters. Thus, although K-Means seems simple enough, it actually works very well

- All_Features_Clustering:
  - I've used the same concepts and applied it to ALL the features
  - Then, using pearson coefficient, I've calculated the correlation between the features. The top two were (Milk, Grocery) and (Grocery, Detergent_Paper). So I've removed Grocery variable and rerun the algorithms again

Only scatter plots are used because using that felt natural and was easily giving insight how well the algorithms were performing. All the plots have zooming ability and all the plots are linked. So by selecting few interesting points, one can readily see how different algorithms are functioning. Many commands which generate .html files are commented, because they are slowing down the program and causing the browser to crash. So, use the .html files directly present in the zip folder

It uses dataset from UCI: Wholesale customers dataset

References:

1. Silhouette Method
2. Pearson Coefficient
3. Bokeh
4. Scikit-Learn
5. Numpy