Application of Clustering Models on Diabetes 130-US Hospitals (1999-2008) Dataset.

By: Rithy Techavoan Yean

Date: 12/1/2024

Submitted to: Dr. Seungjoon Lee for STAT 576 Data Informatics

Motivation

Diabetes is a group of chronic metabolic disorders that affect how our bodies process blood sugar, otherwise known as glucose. Glucose is a source of energy for our bodies and a particular hormone called insulin helps regulate our glucose levels by making sure that our cells uptake glucose. However, insulin can get impaired which means it is difficult to regulate our glucose levels which leads to several health-related problems.

According to the World Health Organization, more than 400 million people in the world are living with diabetes and this number of diabetic individuals is expected to grow to 700 million people by the year 2045 (WHO, 2024). Type-2 diabetes is the most common diabetes which is often linked to dietary habits, exercise frequencies, and age. However, Type-2 diabetes can be preventable if an individual takes the necessary steps to improve their lifestyles. As someone whose family members have type-2 diabetes, I believe that research into this topic to identify patterns associated with diabetes can help diabetes treatment as a whole.

Clustering is an unsupervised machine learning algorithm that groups data points based on their features which can be particularly important in this space. For instance, clustering labels enable the grouping of individuals into their distinct clusters based on their shared patterns in their health data. K-Means clustering for example aims to discover inherent structures or groupings within a dataset. While in the real world, labels are not provided, our investigation can be considered more of a semi-supervised machine learning approach where we can compare clusters with their true labels to see how good the cluster performance is.

Data Summary

The “Diabetes 130-US Hospitals for Years 1999-2008” published by Beata Shack, Jonathan P.DeShazo, and many other contributing authors is a dataset found on the UCI Machine Learning Repository. The dataset is multivariate with both categorical and continuous features. There are 101766 observations with 47 features and each observation represents hospitalized patient records diagnosed with diabetes. The 47 features are explained in detail in Appendix I. The dataset in particular encourages both the use of clustering and classification to draw meaningful conclusions. The target variable ‘readmission’ is about whether a person gets readmitted to hospital for diabetes-related problems upon being discharged beforehand.

The target variable is **‘readmitted’** which represents the following conditions:

1. No: This means that the patient was not readmitted to the hospital within a certain period after discharge.
2. >30: This means that the patient was readmitted more than 30 days after discharge.

3. <30: This means that the patient was readmitted within 30 days after discharge.

Hence, by grouping based on these labels, we can identify shared patterns and key characteristics of patients that place them in each particular readmitted group. Clustering will be useful to see how this happens and to show potentially what features could influence a person being readmitted to hospital or not being readmitted after a certain number of days.

Final Model Description and Result

In the data preprocessing stage, the 50 features were reduced to 36 features. Firstly, features with missing values were removed to avoid generating synthetic data based on median and mode imputation. Features with only one value were removed due to their redundancy. Identifying features like a patient's ID, discharge ID, admission type ID, and patient number were all removed based on redundancy as well.

Dimensionality Reduction

The purpose of dimensionality reduction is to avoid the curse of dimensionality which is where data becomes sparse and therefore becomes much harder to cluster. Through feature selection, only the relevant features are selected which makes clustering much more interpretable and computationally efficient.

Correlation Matrix

The first dimensionality reduction technique used was to check the correlation of the numerical features in the dataset. However, the correlation of the numerical features did not exceed 0.5 and so removing features by this method did not seem robust. This is because a correlation of 0.5 or less suggests that the features are not highly correlated at all. The goal of correlation matrix feature selection is to remove highly correlated features and therefore this method was not appropriate for this particular dataset.

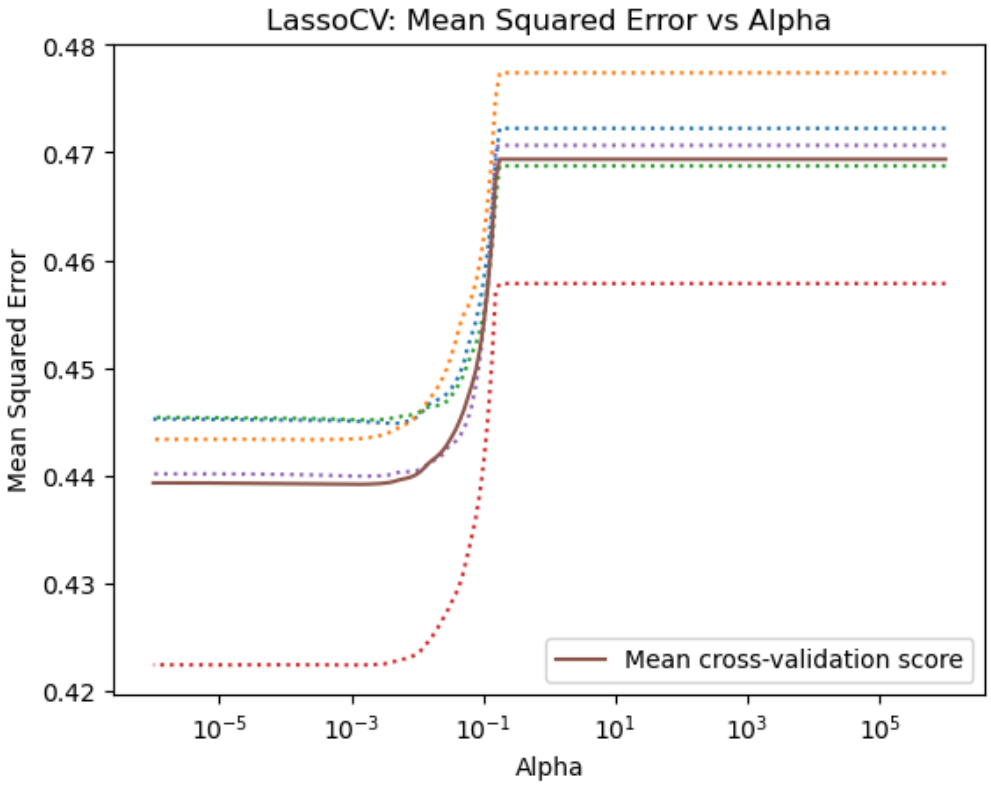
Lasso Regression (L1 Regularization)

Since the dataset have both numerical features and categorical features, and the categorical have an ordinal nature about them, label encoder was used to assign each categorical variable a unique integer value. From there, Lasso Regression was used to select the most relevant features. The process of Lasso Regression is as follows:

$$\sum_{i=1}^{n}(y_i - \sum_j (x_{ij}\beta_j)^2 + \alpha \sum_{j=1}^{p} |\beta_j|$$

The Residual of Sum Squares (RSS) measures how well the model fits the data. This is presented by the following $\sum_{i=1}^{n}(y_i - \sum_j (x_{ij}\beta_j)^2$. The α is the regularization parameter that controls the strength of the penalty which in turn determines how much shrinkage occurs. β_j are the model coefficients and the L1 penalty $\sum_{j=1}^{p} |\beta_j|$ encourages the model to have sparsity as some coefficients are shrunk which thereby reduces the dimensionality of the model.

The key idea of L1 regularization is to penalize the absolute values of the coefficients. The coefficients that are shrunk to zero are deemed as less important features and the coefficients that are non-zero are deemed as the important features. For this particular investigation, the choice of alpha is important as a higher alpha value could leave less features to work with. The optimal value of alpha was chosen by using cross-validation.



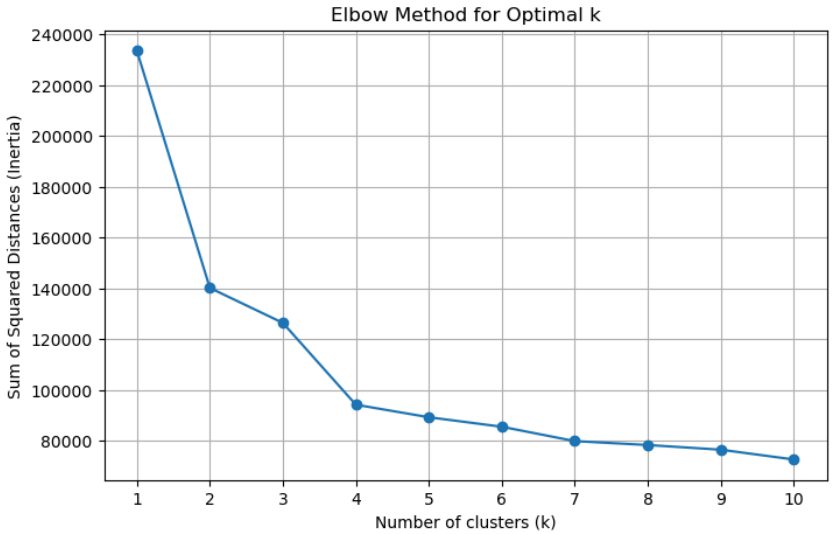
The ideal choice for alpha is the elbow point by looking at the mean cross-validation score curve. Using this curve, it appears that alpha is roughly 0.01 or 10^{-2}. Using this alpha value, the relevant features were selected as follows:

From Lasso Regression, the following 11 features were selected:
Age, Admission_source_id, Time_in_hospital, Num_lab_procedures, Num_procedures, Num_medications, Number_Outpatient, Number_emergency, Number_inpatient, Number_diagnoses, and DiabetesMed.

Instead of working with a dataframe of (101766, 36), the dimension was reduced to (101766, 11). Moreover, while it is ideal to run the different clustering algorithms on the entire observations of the dataset, the downside is the computational complexity. One method to handle such computational complexity was to sample only 10% of the entire dataset. This was done using the resample function of the scikit-learn utilities package. Hence, the sampled data used for clustering had a dimension of (10176, 11) which eased computational complexities.

Lasso Regression and Feature Selection was very much needed as the default model with all the features involved had a silhouette score of only 0.203 and an ARI index of -0.0079 which suggests that the clustering is not well-defined and that the points themselves were not clustered to their true labels.

K-Means Clustering (Default Model)



As part of the hyperparameter tuning process, the optimal K-Means clustering must have the highest silhouette score and the elbow method was used to determine that ideal number of clusters. It is clear that n_cluster = 2 is the optimal number of clusters for this particular dataset.

At this optimal number of clusters, the silhouette score is reported as 0.392. Any other combination of clusters will result in a lower silhouette score.

The silhouette score is a metric used to evaluate the quality of clustering results by measuring how similar an object is to its cluster, also known as cohesion, against other clusters, known as separation. The silhouette score is defined as follows:

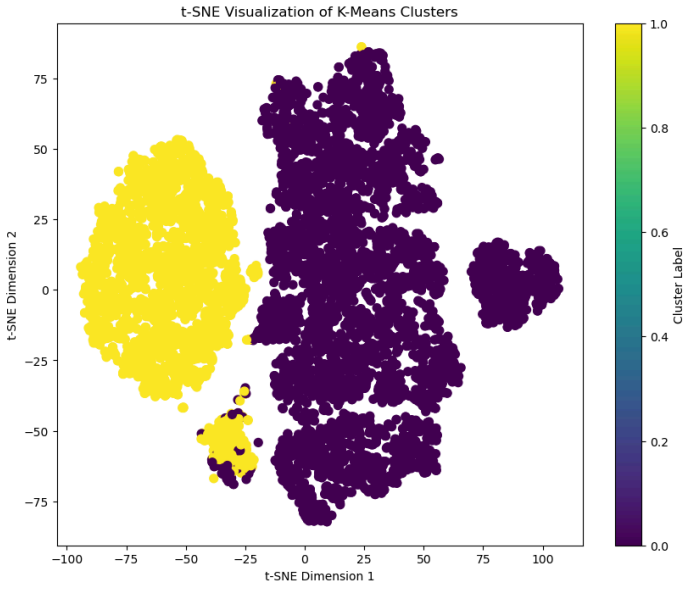
$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

Where b(i) represents cohesion and a(i) represents the separation. A score of 0.392 suggests that the clustering has moderate clustering ability. Since the score is positive, the points clustered are somewhat closer to other points in the same cluster instead of closed to other clusters. While the clusters are reasonably separated, they are not highly distinct which suggests that there could be room for improvement.

Moreover, the adjusted rand index (ARI) was used to evaluate how closely the clusters aligned with ground truth labels. The ARI index is calculated as follows:

$$ARI = \frac{correct\ similar\ pairs + correct\ dissimilar\ pairs}{total\ number\ of\ pairs}$$

The ARI index is -0.0068 for the default K-Means Model which suggests that it is worse than random meaning that points in the same cluster are actually being placed in opposite clusters instead. Therefore, other clustering algorithms such as DBSCAN and Agglomerative Clustering can be used to capture the data especially since the data is unlikely to be linear.



Moreover, the default model only separated the features into two clusters as seen in the t-SNE visualization. This is because it believes the optimal number of clusters is 2 through the elbow method which maximizes the silhouette score. Within the context of our dataset, this is not a good clustering because we have three different target labels. Hence, there should be three clusters for the points to be allocated to their appropriate groupings. This default model showcases how there should be more clusters to group data points more accurately based on their shared characteristics. Hence, more exploration and more model combinations are needed.

Comparison Table of Different Models

To compare the different models, the silhouette score and ARI score will be used. In addition to looking at these silhouette scores and ARI metrics, the clustering methods were also visualized in 2D and 3D space. This is because visualization enables us to see how well-separated the clusters are and if they overlap or not. It also gives good indication of the noise points in the dataset and as well as how well-defined each cluster can be, This is not apparent through the silhouette score and ARI metric. Hence, visualization is key to choosing the best clustering as well. For visualization, MDS was attempted but long computational complexity led to Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) taking precedence for visualization. The visualizations graphs are provided in Appendix II.

Table 1: Comparison of different models of clustering and their silhouette and ari scores

Model:	Parameters:	Silhouette Score:	ARI:
Default Model KMeans (no feature selection)	optimal n_cluster = 2	0.203	-0.0079
Default Model KMeans (w/ lasso feature selection)	optimal n_cluster = 2	0.392	-0.0068
DBSCAN (default)	eps = 0.5, min_samples = 5	-0.521	-0.0019
DBSCAN (optimal)	eps = 2 min_samples = 10	0.327	0.018

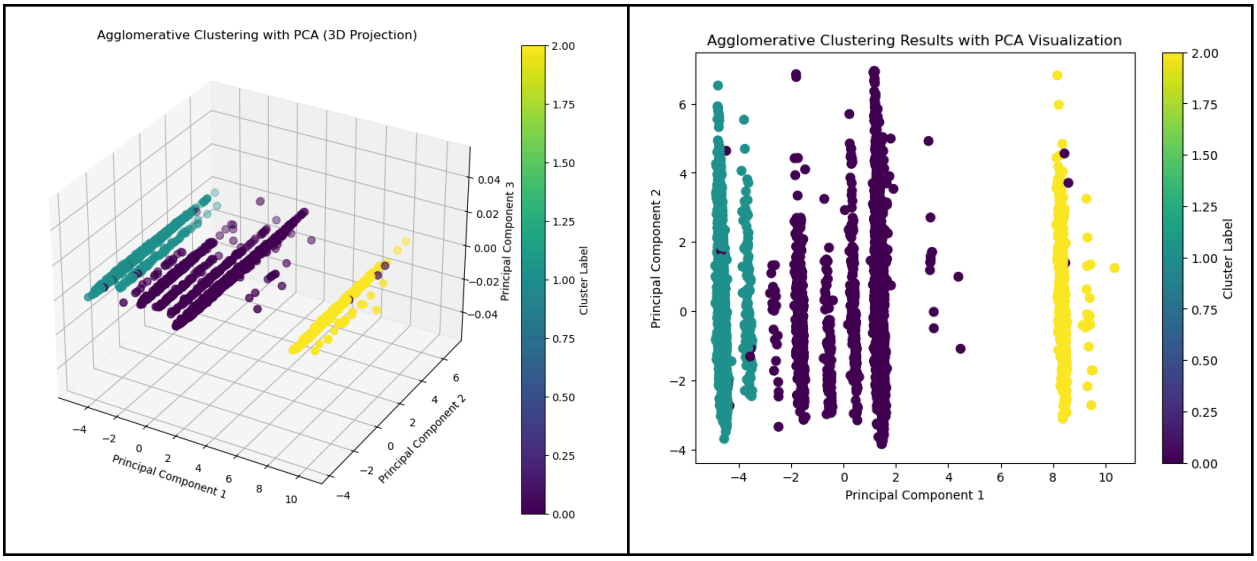
Agglomerative Clustering (optimal)	n_cluster = 3 linkage = ‘ward’	0.422	-0.00736
Agglomerative Clustering	n_cluster = 3 linkage = ‘complete’	0.631	0.00125
Agglomerative Clustering (optimal)	n_cluster = 2 linkage = ‘complete’	0.633	0.001
Agglomerative Clustering	n_cluster = 3 linkage = ‘average’	0.570	0.00269
Agglomerative Clustering	n_cluster = 3 linkage = ‘single’	0.620	0.00010
Agglomerative Clustering (optimal)	n_cluster = 10 linkage = ‘single’	0.792	0.001

Discussion and Analysis

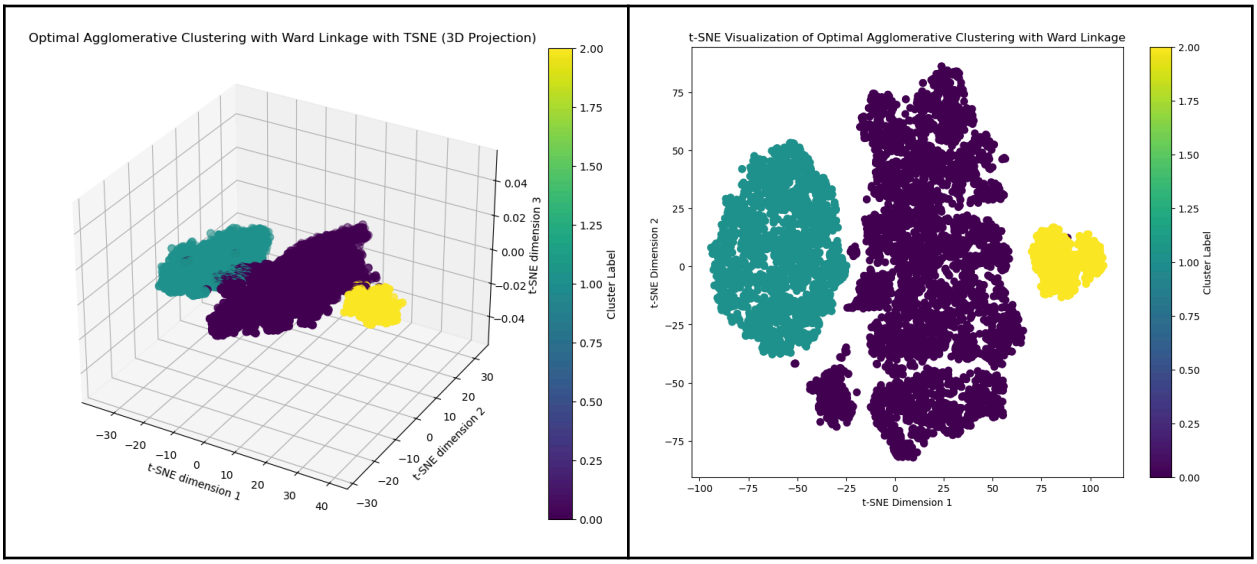
The different model combinations were decided based on default metrics and hyperparameter tuning. Normally, prioritizing the silhouette score would mean that Agglomerative Clustering with the optimal parameters with a score of 0.792 would be chosen. However, there are limitations to just looking at the silhouette score as it has limitations for non-spherical and overlapping clusters. For instance, a high silhouette score is achieved but the clusters themselves are not distinct enough especially if the data contains significant noise and if the clusters are irregular in shape which is apparent for this dataset. Using this same reasoning, other model combinations with high silhouette score such as Agglomerative Clustering with two clusters and with complete linkage method (score: 0.633) and Agglomerative Clustering with three clusters and with average linkage method (score: 0.620) were dismissed because their clustering visualizations did not separate the data points well enough. The DBSCAN default model had a silhouette score of (-0.521) which suggests that the clustering structure of the data is very poor which basically implies that outliers that do not belong to any cluster are being assigned to the clusters. This is why there was a need to optimize the DBSCAN model which led to a silhouette score 0.327 which was still lower but much better than the default model. However, looking at the diagram of the new DBSCAN model in Appendix II, we could see that there were many overlaps of the clusters which suggest very poor clustering of the target.

Looking at all the figures of each of the clustering from Appendix II, the best clustering visually is the Agglomerative Clustering with optimal parameters n_cluster = 3 and linkage = ‘ward’ which has a low silhouette score of 0.422 but has good clustering visually.

PCA Visualization with Three PCA Components and Two PCA Components:

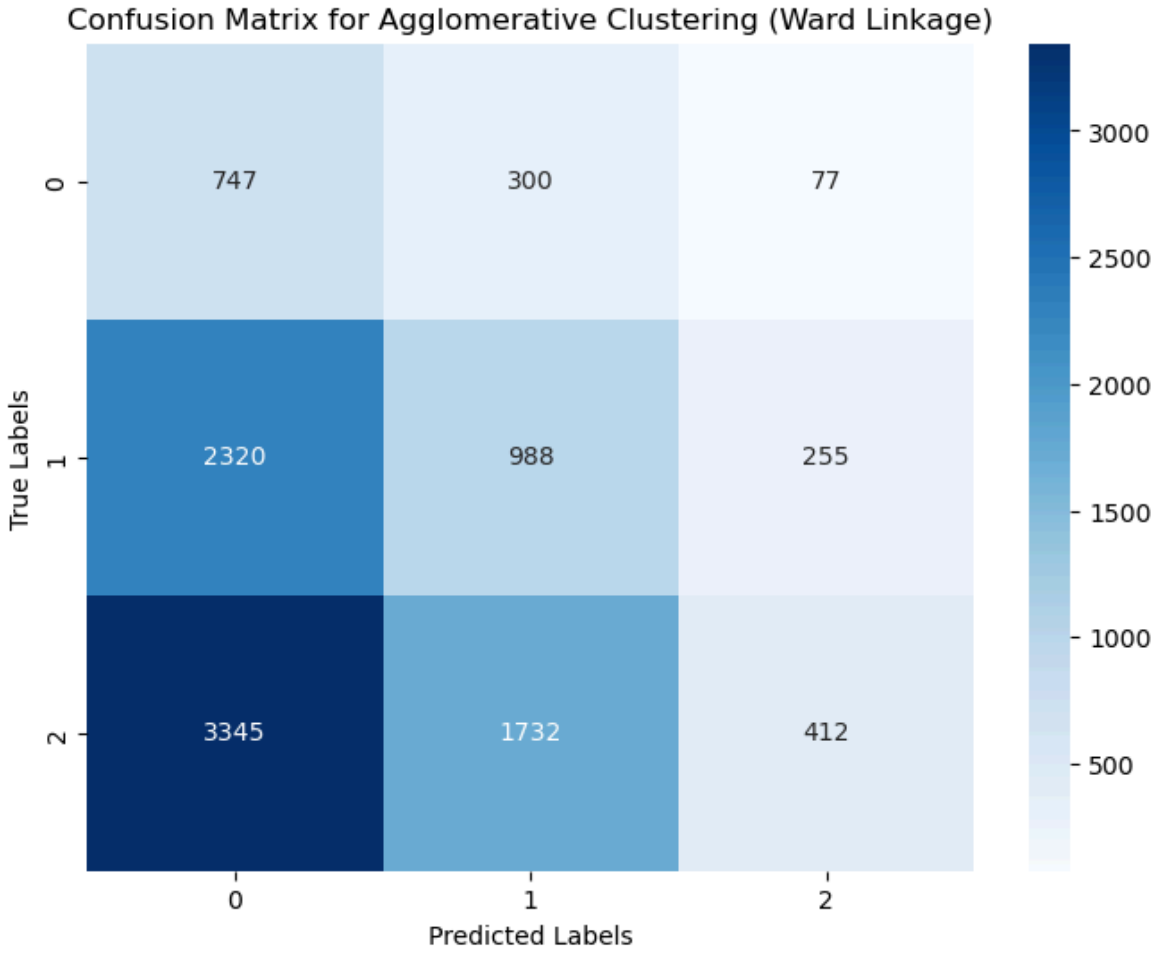


t-SNE visualization with Two Dimensions and Three Dimensions Respectively.



Looking at the visualization of the clusters, it is shown that for the three labels of (0, 1, 2) where 0 represents No readmission, 1 represents >30 (readmitted after more than 30 days), and 2 represents <30 (readmitted after less than 30 days). It is clear that the clusters are much more

well-defined than the other clusters based in Appendix II. Keep in mind that these clusters are the result of the features being transformed into a lower dimensional space. The t-SNE visualization is more defined as the yellow cluster is well-separated from the blue and purple clusters. While there are a few purple data points that are very close to the blue cluster, they are not classified into the blue cluster. One potential indicator for improvement for the dataset is perhaps to introduce another new label. For instance, there is a smaller purple cluster in the 2D t-SNE visualization which could be its own separate cluster as it is farther than the larger purple cluster in the center.



From the confusion matrix, we can see that the true labels and the clustering are not matching well which suggests why the ARI score is quite low. Suggesting that the clusters, while well-defined, are not great at allocating the appropriate data points to their appropriate labels. This clustering was the best in terms of confusion matrix as well. Therefore, there were many wrong classifications if we consider the true labels of the dataset. In the real-world, of course we

do not know what these labels are, so wrongful classifications could be costly in the context of health.

Within the domain of diabetes, the findings in this investigation does suggest that certain features that were selected for clustering such as Age, Admission_source_id, Time_in_hospital, Num_lab_procedures, Num_procedures, Num_medications, Number_Outpatient, Number_emergency, Number_inpatient, Number_diagnoses, and DiabetesMed could influence a person being readmitted into hospital for diabetes-related problems. For instance, it's clear that based on the number of clusters, there are certain indicators for why a person may be classified into not being readmitted, readmitted after more than 30 days, and readmitted after less than 30 days. However, the investigation does not conclusively say which features are responsible as that is not within the scope of this exploration.

The limitation of clustering is that it does not tell which feature is directly responsible for this as the clustering was done using PCA and t-SNE. PCA and t-SNE reduce the dimensionality of a data in a lower-dimensional space which can help make the data visually interpretable but it comes with its own set of challenges. For instance, it does not preserve the true structure of the original dataset which has much higher dimensions. PCA prioritizes variance which can lead it to remove subtle structures that are important for clustering but do not contribute to variance. Whereas, t-SNE emphasizes the local structure but distorts the overall global structure. Moreover, the dimensionality reduced structures are harder to relate back to the original feature set. The clusters may appear distinct but they are represented in terms of transformed dimensions which again cannot be interpreted as the original features. Therefore, the features selected through feature selection for this dataset cannot directly be attributed to the distinct clusters shown in the visualization as they are extractions of the original features represented in a lower-dimensional space.

For future exploration, both the silhouette score and ARI index on top of the visualizations should be optimized for the best clustering outcome. Due to the time constraints of this project, this was not achieved. However, it serves as a good start in terms of exploring clustering possibilities with this dataset which can help provide useful information for hospitalization readmission for diabetes.

Appendix I:

Explanation of Features within the Dataset

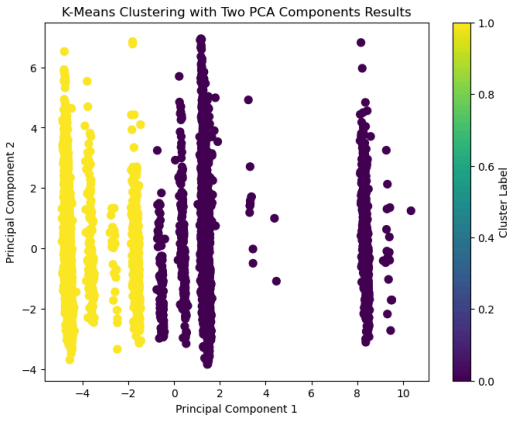
Feature Name	Data Type:	Explanation of Feature:
encounter_id	**Unique Identifier (integer)**	The unique ID for each patient in the hospital.
patient_nbr	**Unique Identifier (integer)**	A unique ID for each patient in the dataset.
race	**Categorical** White, African American, Asian, Hispanic, Other	The race of the patient.
gender	**Categorical** Male, Female, Unknown/Invalid	The gender of the patient.
age	**Categorical** [0-10), [10-20), [20-30), [30-40), [40-50), [50-60), [60-70), [70-80), [80-90), [90-100)	The age group of the patient.
weight	**Categorical** Missing, Normal, Overweight, Obese	Weight category of the patient.
admission_type_id	**Unique identifier (integer)**	The type of admission. It indicates the nature of the hospital admission.
discharge_position_id	**Categorical** Discharged, Transferred, Died, or Others	The status of the patient at discharge.
admission_source_id	**Categorical** Referral, Direct Admission, Emergency Room, Other	The source of admission which indicates how the patient entered the hospital.
time_in_hospital	**Numerical**	The number of days the patient stayed at the hospital
num_lab_procedures	**Numerical**	The number of lab procedures the patient underwent during their hospital stay
num_procedures	**Numerical**	The number of procedures the

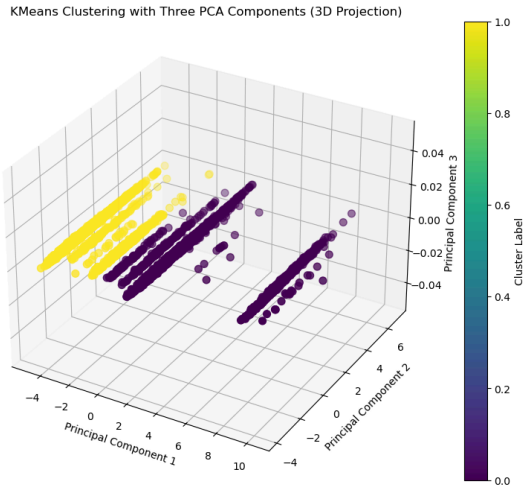
		patient underwent during their hospital stay
num_medications	**Numerical**	The number of medications the patient was prescribed during their hospital stay
number_outpatient	**Numerical**	The number of outpatient visits a patient had prior to the hospitalization
number_emergency	**Numerical**	The number of emergency visits a patient had prior to the hospitalization
number_inpatient	**Numerical**	The number of inpatient visits a patient had prior to the hospitalization
diag_1, diag_2, diag_3	**Categorical**	Diagnosis codes for the patient's primary, secondary, and tertiary codes (ICD-9)
num_diagnoses	**Numerical**	The total number of diagnoses that patient received during their hospitalization
max_glu_serum	**Categorical** Norm (Normal) 200 (Above 200) 300 (Above 300) None (missing)	The maximum glucose serum level
A1Cresult	**Categorical** Norm (Normal) 7 (Above 7) 8 (Above 8) None (missing)	A1C test result (used to measure long term glucose control)
metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, glipizide, glyburide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone,	**Binary** **(0-1)** 1 means the patient was prescribed the medicine and 0	These features indicate whether the patient was prescribed any of the following medicines during their hospital stay.

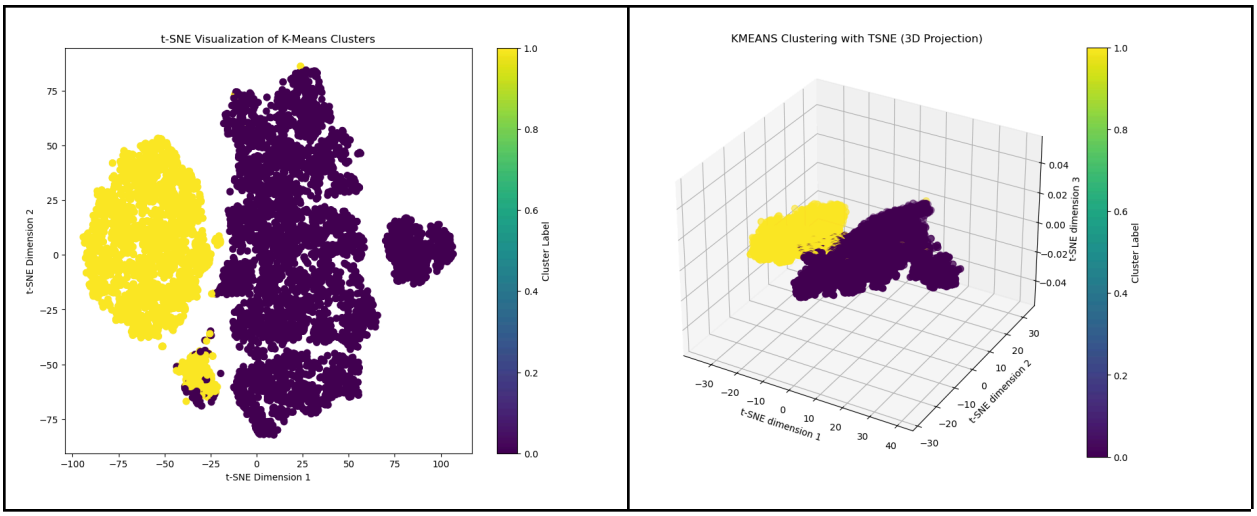
tolbutamide, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone:	means they were not.	
change	**Categorical** Increased, decreased, or unchanged	Indicates whether there was a change in patient's medication during their hospitalization.
diabetesMed	**Binary** **(0-1)** 1 means they were prescribed diabetes medication and 0 means they were not	Whether a patient was prescribed a diabetes medication
readmitted	**Categorical (Target Variable)** No (patient was not readmitted) >30 patient was readmitted after more than 30 days <30 patient was readmitted after less than 30 days	Whether the patient was readmitted to the hospital within 30 days of discharge.

Appendix II:

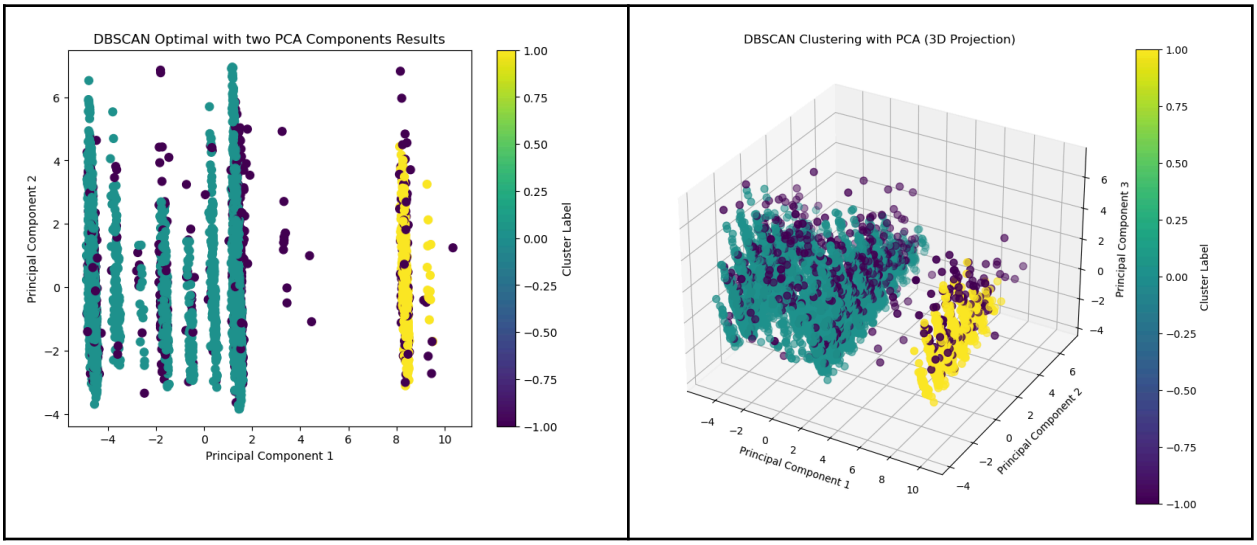
K-Means Clustering Visualizations

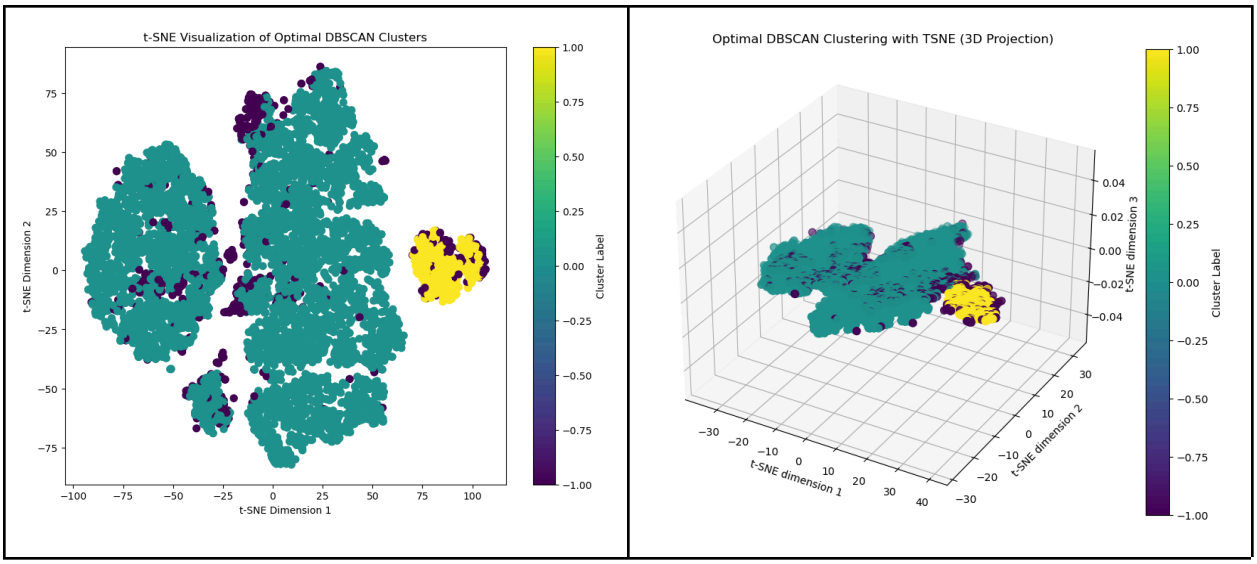




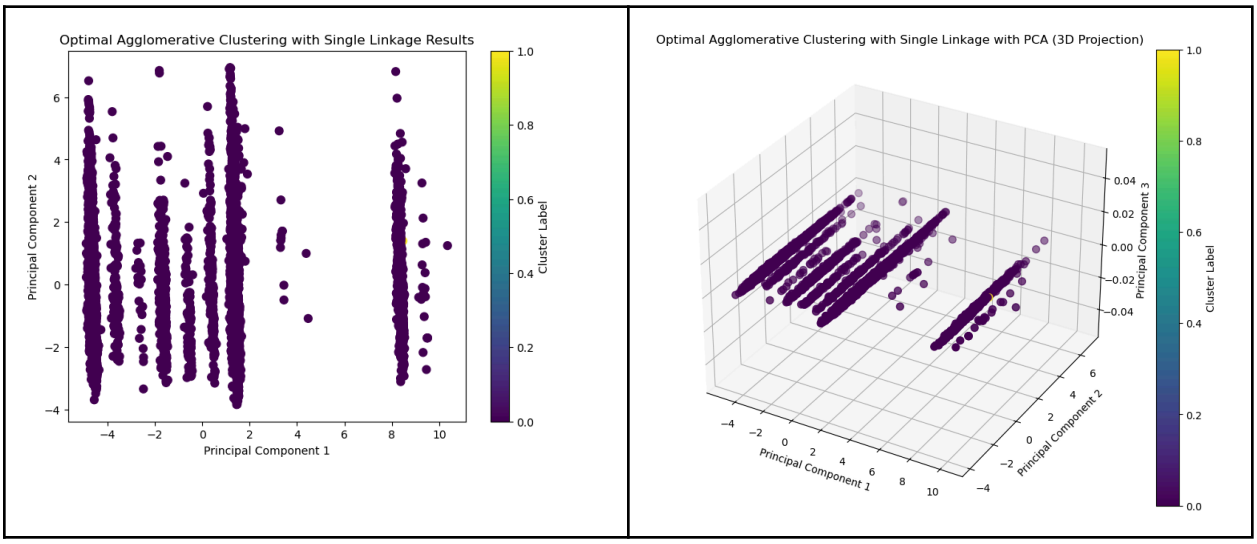


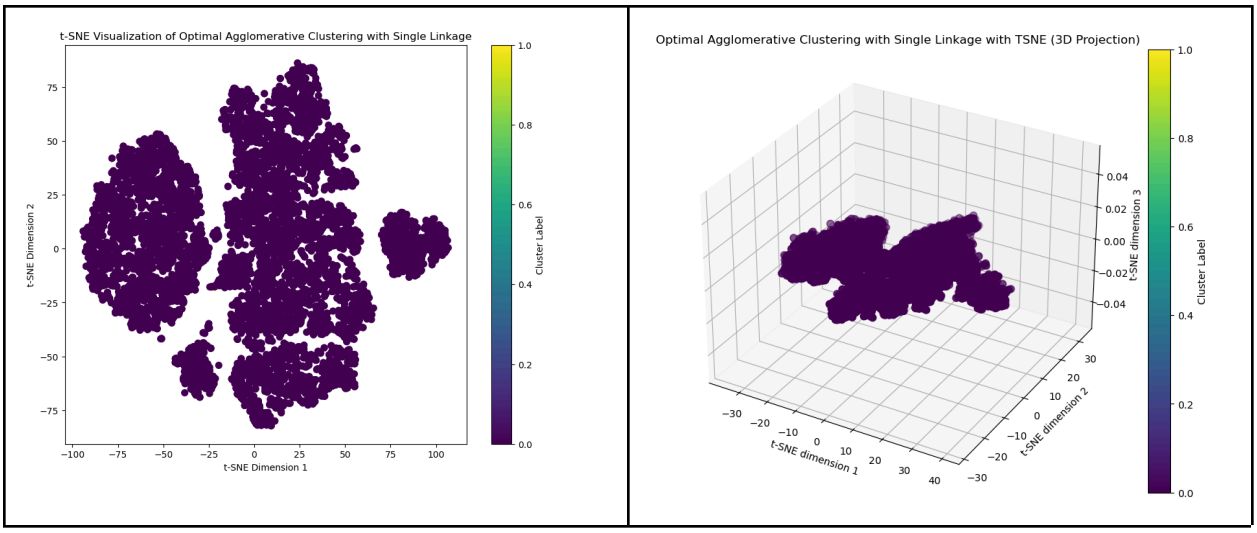
DBSCAN Clustering Visualizations



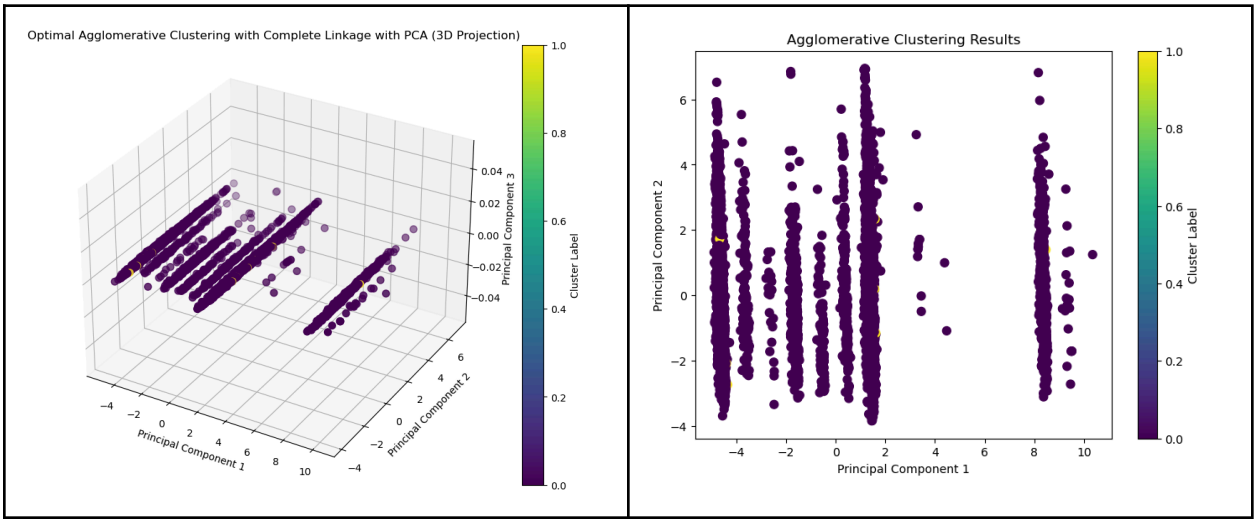


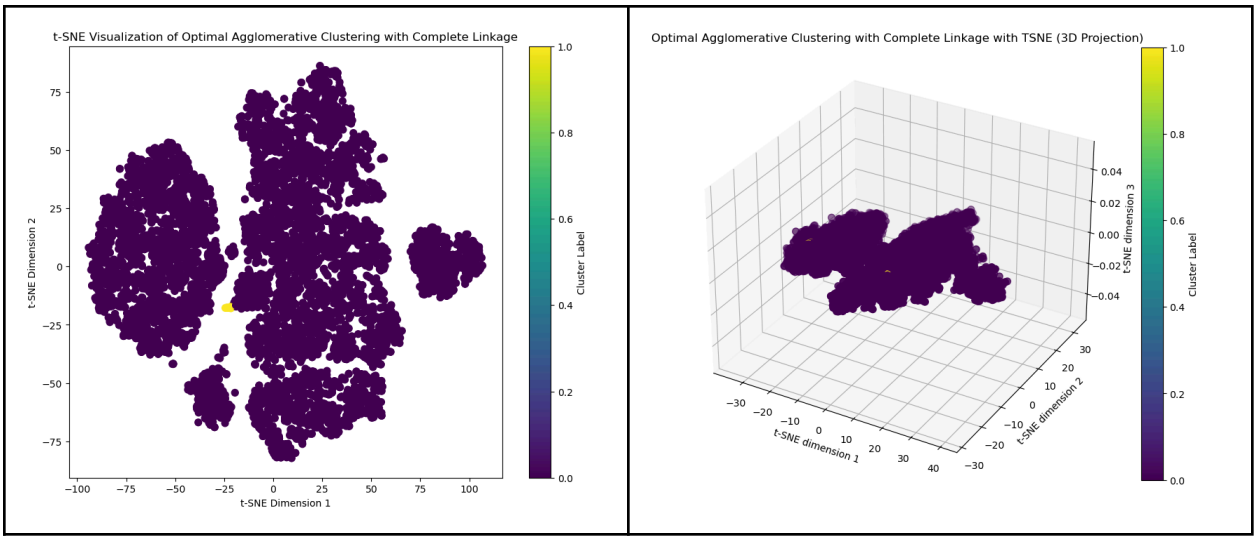
Agglomerative Clustering: Single Linkage Visualizations

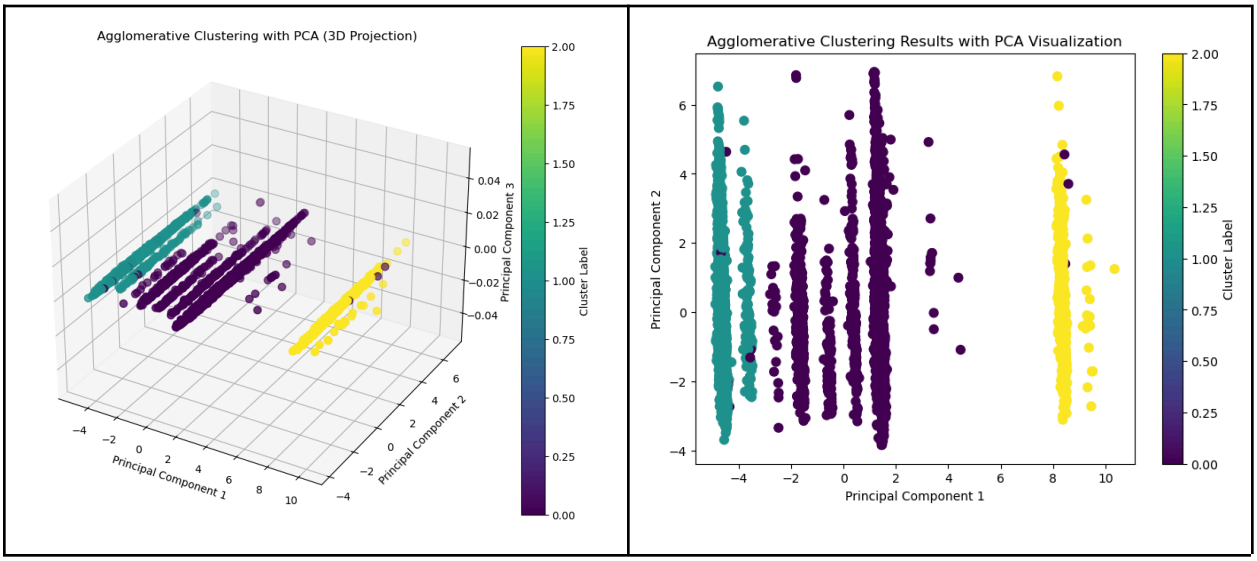




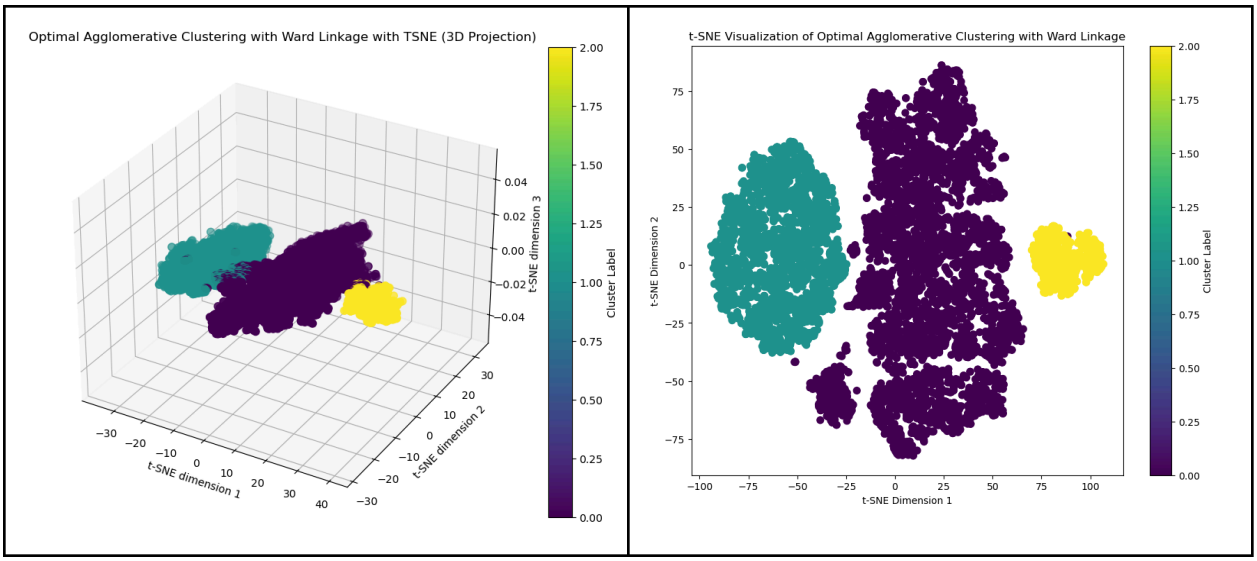
Agglomerative Clustering: Complete Linkage Visualizations







t-SNE visualization with Two Dimensions and Three Dimensions Respectively.


Optimal Agglomerative Clustering with Ward Linkage with TSNE (3D Projection)
t-SNE dimension 1
t-SNE dimension 2
t-SNE dimension 3
Cluster Label
t-SNE Visualization of Optimal Agglomerative Clustering with Ward Linkage
t-SNE Dimension 1
t-SNE Dimension 2
Cluster Label