Unsupervised Learning: A3 - CS7641

Perry Francois-Edwards (pdfe3@gatech.edu)

1 INTRODUCTION

This report will investigate the performance and functionality of clustering and dimensionality reduction on datasets and neural networks. When comparing the algorithms, the results will analyze metrics, methods, learning curves and clock times.

1.1 Datasets for Clustering and Dimensionality Reduction

The two datasets chosen were ported from A1: the Cars Acceptability Evaluation (CAE) and the Red Wine Quality Measure (RWQM). Though CAE yielded higher f1_scores when running supervised learning algorithms, I expect that its categorical features will cause issues when running the unsupervised learning algorithms, even with the use of one-hot encoding. Thankfully, RWQM has continuous values, which should perform better within the clustering and dimensionality reduction methods. One drawback to both these datasets is that they have relatively low number of features; (21) for CAE and (11) for RWQM. That being said, it opens the possibility for interesting dimensionality reduction results. An additional difference in A3, the CAE dataset will use binary onehot encoding instead of its previous numerical encoding. Here, the distance variables should provide more information with increased features.

1.2 Hypothesis

RWQM will provide clear and distinct values for clustering to prove the optimal number because of its ability to determine distances.CAE will show a larger spread of viable cluster values, due to its categorical design.

After viewing the RWQM dataset, I hypothesize that a few of its features are not important and Dimensionality Reduction will work effectively on this dataset. However, with the new structure of CAE, I hypothesize that Dimensionality Reduction will be able to describe the characteristics of the data structure, but it won't be as effective or definitive as RWQM.

Since the Dimensionality Reduction datasets are expected to reduce both datasets, it can be hypothesized that the applied clustering algorithm will redefine and reduce the optimal clusters, due to a reduction of dimensions.

With dimensionality reduction and its method usage, when using the dataset on the Neural Network (NN), the convergence and performance will be worse due to a loss (in some respect) of information, dependent on the dimensionality reduction algorithm. The loss is anticipated to be greater than the baseline.

With Clustering algorithms, the performance of the NN on the dataset is hypothesized to improve if the added clustering feature correlates to the target values, thus having clear and defined clustering metrics. However, the fit times are expected to take longer.

2 METHODS

2.1 Clustering and Dimensionality Reduction (Steps 1 to 3)

Within this experiment, clustering and dimensionality reduction will be used on the training data from the 80/20 split.

For step 1, two clustering algorithms, Expectation-Maximization (EM) and K-Means, will be used on both CAE and RWQM. To obtain the optimal number of clusters for EM, the

Bayesian Information Criterion (BIC), an unsupervised metric, was used to test the model and the optimal clustering value is the minimal BIC. To obtain the optimal number of clusters for the K-Means, the Within-Cluster Sum of Squares (WCSS) metric was used to find the optimal number of clusters by the elbow method.

For Step 2, the dimensionality reduction algorithms, Principal Component Analysis (PCA), Independent Component Analysis (ICA), Random Components Analysis / Random Projection (RCA) and Isometric Mapping (Isomap), will be applied. To choose the optimal features, PCA uses a Scree graph that shows the variance explained by component and the components were selected based on the Kaiser's rule and elbow point method [5]. ICA chose its optimal values by reviewing kurtosis and selecting the maximum values above a threshold. RCA uses reconstruction error to find the optimal component values by using the value with the lowest error or knee point. The Isomap also uses reconstruction error to find the feature count determined by the elbow method.

For Step 3, the clustering algorithms were applied to the newly transformed datasets from Step 2. The same metrics and methods were used to determine the optimal clusters.

2.2 Neural Networks (Steps 4 & 5)

The RWQM dataset and the NN algorithm process from A1 were used to test how well the clustering and dimensionality reduction converged to the f1 score. Similar to A1, the training data will have a random 80/20 split.

For Step 4, the Neural Network algorithm was performed four separate times with the dimensionality reduction algorithms (PCA, ICA, RCA and Isomap) transformed data based on optimal feature reduction and those four NN outputs were compared to the original Neural Network output. The ICA and Isomap algorithms will be reported. The learning curves convergence and loss curves will be compared.

For Step 5, the same dataset will be used as step 4, but will be testing the clustering algorithms. The two outputs of the clustering methods will be added as one column feature into the dataset. The output dataset will then be used with the NN and compared to the baseline NN based on learning curves convergence and clock times.

3 CLUSTERING (STEP 1)

The scikit-learn, mixture.GaussianMixture and cluster.KMeans algorithms with default settings, were utilized. Figure 1 displays the clustering means algorithms on the RQWM and CAE datasets.

3.1 Results

Since the RWQM clusters are similar for both algorithms, it reconfirms its validity. It's quite interesting that the RWQM cluster amount is closely aligned with the number of labels in the target class and it makes sense. The cluster should group values together that align well. The clusters have similar unique value imbalances that align with the original datasets target values.

The CAE has varying outcomes of clusters, even though its low value of BIC does not vary much over a wider span of cluster values, displaying variability. Since the original data was categorical, it makes sense that clustering distances do not clearly outcome to a specific cluster for similar ranges. These clusters do not align with the labels in the target class due to the vast majority of data.

Overall, the performance was directly correlated to the data structure. However, the clear algorithmic winner in this comparison is the influence of a stronger and more complex

algorithm in EM and it works better on data that has mixed distributions. Since neither dataset has clear separation boundaries, EM is expected to work best. To improve outcomes, removing outliers in the preprocessing steps could have helped these algorithms perform better since that is a weakness of clustering.

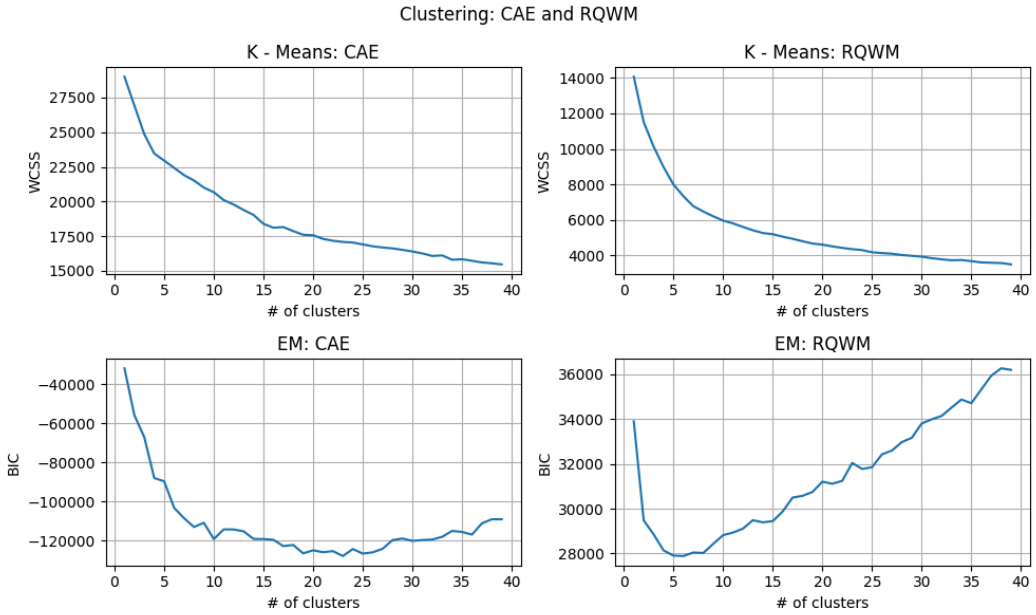

Figure 1—K-Means and EM metrics graphs for both datasets.

3.2 KMeans

Figure 1 showcases the metric to choose the optimal number of clusters for K-Means. The Car dataset does have a clear elbow point, but the conservative choice would be cluster=15, but it could also be closer to 4. This example shows the simplistic nature of K-Means and the drastic variability it has when selecting clusters. The RWQM dataset has a clear cluster number, cluster=6.

3.3 EM

Figure 1 showcases EM's variable BIC. The results are expected as the number of clusters to generalize the data for a categorical value dataset of CAE is more than a continuous values dataset. Here the optimal number of clusters for CAE=23 and RWQM=6.

4 DIMENSIONALITY REDUCTION ALGORITHMS (STEP 2)

The scikit-learn, decomposition.PCA, decomposition.FastICA, random_projection and manifold.Isomap algorithms with default settings were utilized.

4.1 Results

The data results for dimensionality reduction all vary, but it seems that besides ICA, which seems to characterize data variation more than performance, the RWQM had a higher ranked dataset than the CAE. The noise seems to affect PCA and ICA similarly for the CAE data because the clustering algorithm reaches its methods threshold at a similar component. Even Isomap reaches its minimum at that point.

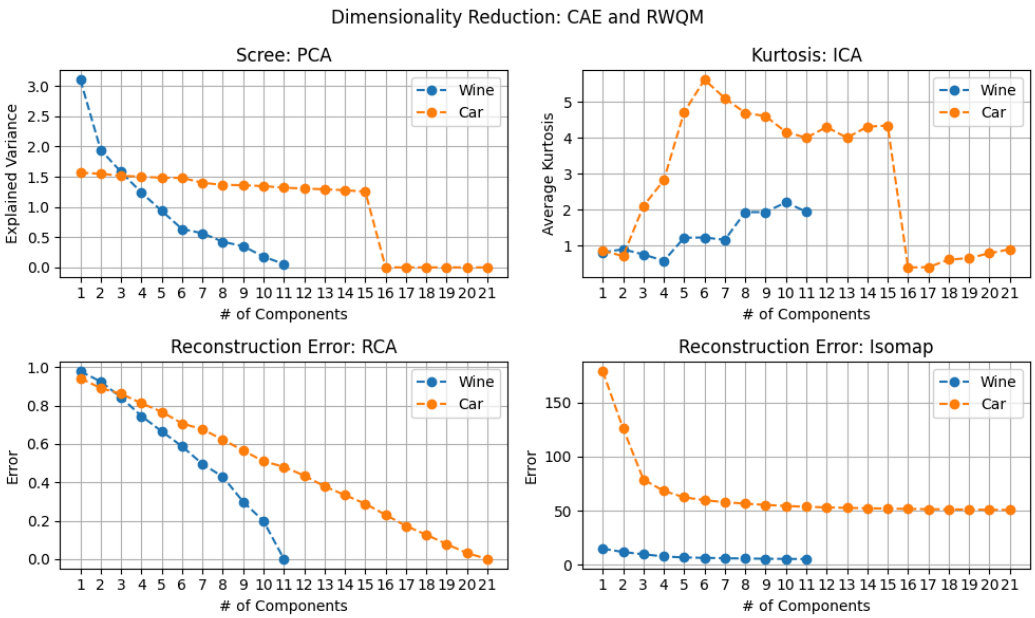

Figure 2—PCA, ICA, RCA and Isomap metrics for both datasets.

4.2 PCA

The eigenvalues for CAE range from slightly above 1.5 to 0, with values greater than 15 all equaling 0. The distribution of eigenvalues for RWQM are slightly above 3 to right above 0. The optimal component will be determined by Kaiser's Rule, which denotes that any value with explained 'variance less than 1 contains less information' in comparison to the larger values[5]. This was then followed by the elbow method. For RWQM, the elbow method could range from 2 to 6, but with Kaiser's Rule, the optimal component is 4. For CAE, Kaiser's Rule and the elbow method align at 15 components. RWQM has a higher collinearity than CAE quantitatively by the Scree plot because its first few components have a higher explained variance. With a scree plot of explained variance, noise would affect the metric by having low variances.

4.3 ICA

The Kurtosis threshold was greater than 1 component. I chose this variable to be consistent between the two datasets and to have a balance of widespread data and possible outliers. The Kurtosis of CAE, depending on the number of components, can be either Leptokurtic (Kurtosis> 3 & components between 5 and 15) or Platykurtic (Kurtosis <3 & components <5 and >15) [3]. All components of RWQM lie in the Platykurtic region. Thus meaning that there are less outliers in this data and that it is more spread out in comparison to its mean[3]. Noise is apparent in the Kurtosis dataset by having the value's drop drastically after 15 components. The CAE data has more independence.

4.4 RCA

Random projection uses reconstruction error to determine the number of components. This error, mean squared error, was the difference between the reconstructed dataset and the original dataset. Unlike Isomap, the reconstruction error map for RCA does not have an indication of feature reduction. Thus, for this scenario it would be best to use the

maximum number of features for the number of components based on Figure 2. Or, one could use the knee method for the RWQM dataset. The more components, random projections, the lower the error which is expected based on current features in the dataset. With the total number of features, the expected and hypothesized error is zero.

4.5 Isomap

Isomap also uses its inherent reconstruction error to determine the number of components to use. However, this use of the error is found by geodesic distances between points [4]. Here, the curve of the error metric can determine the number of components by the elbow method as well. Here, both the CAE and RWQM optimal number of components is 3.

5 CLUSTERING ON THE DIMENSIONALITY REDUCTION ALGORITHMS (STEP 3)

The clustering algorithms were used to explore if the applied Dimensionality Reductions would change the expected cluster outcomes. The focus will be on K-Means clustering on Random Projection and Isomap for both datasets. The CAE dataset is more interesting because of its improved range of variability.

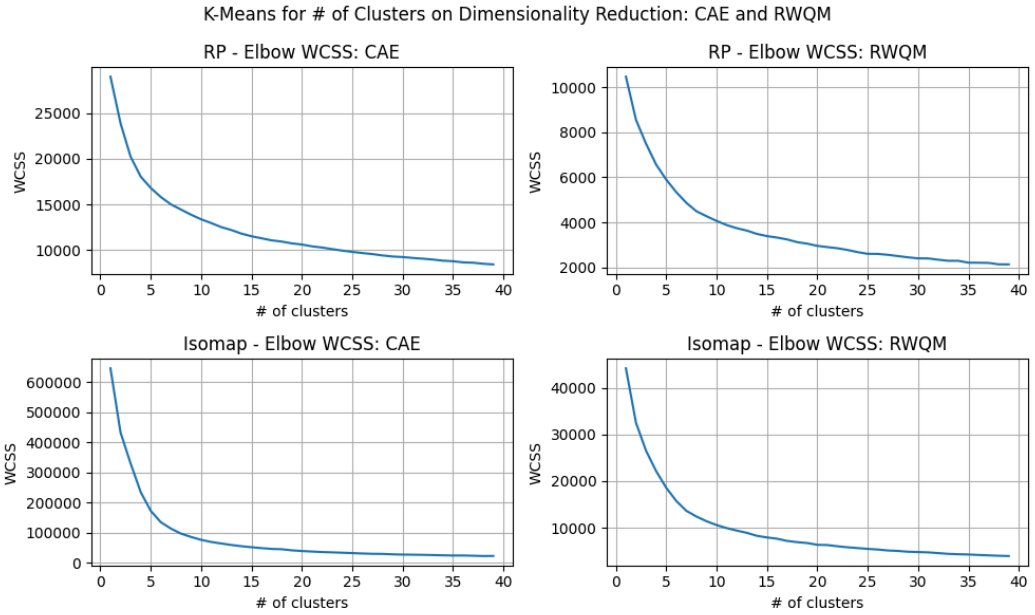

Figure 3—K-Means clustering applied to Random Projection and Isomap algorithms on both datasets.

The RWQM dataset produces similar cluster values, as Step 1, clusters=6, for both dimensionality reduction algorithms. When looking at Figure 3, the Isomap and Random Projection algorithms applied to the CAE datasets seem to have more defined elbows here in comparison to Step 1. Though the gradient seems to vary, the cluster values seem to both align at clusters=4. Here, the clusters do not leave much room for interpretation. After reducing the dimensions, the algorithm is also reducing the curse of dimensionality which can help clustering dimensions increase information and meaning.

6 NEURAL NETWORK (STEP 4 & 5)

The scikit-learn neural networks, MLPClassifier, was used along with activation='relu' and solver='adam'. The optimal components and clusters were derived from steps 2 and 3. The RWQM dataset will be investigated because of its continuous variable structure.

6.1 Dimensionality Reduction Algorithms

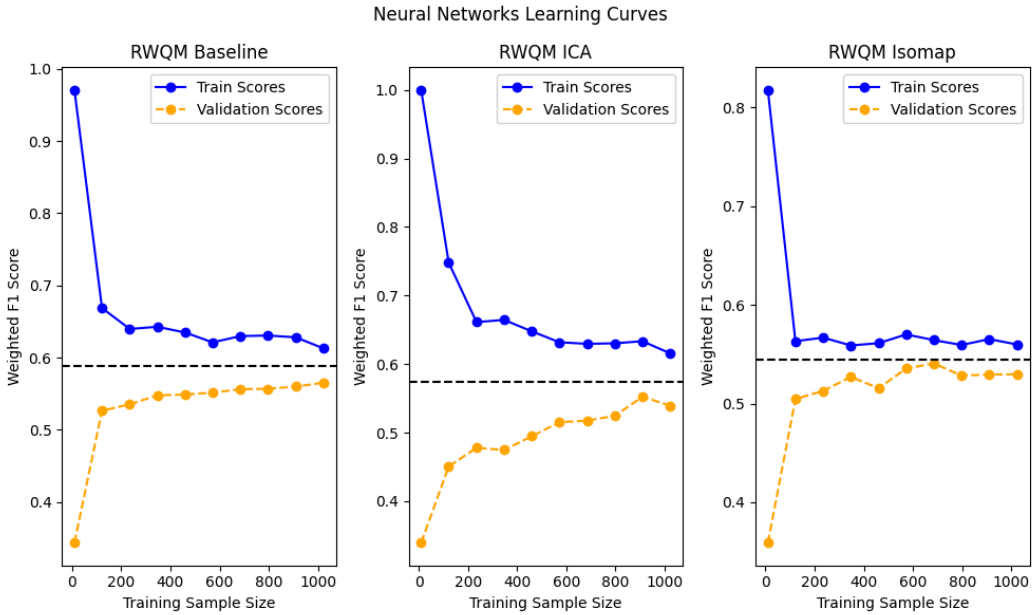

Figure 4—The Baseline (A1), ICA and Isomap Learning Curves with optimal algorithm components applied.

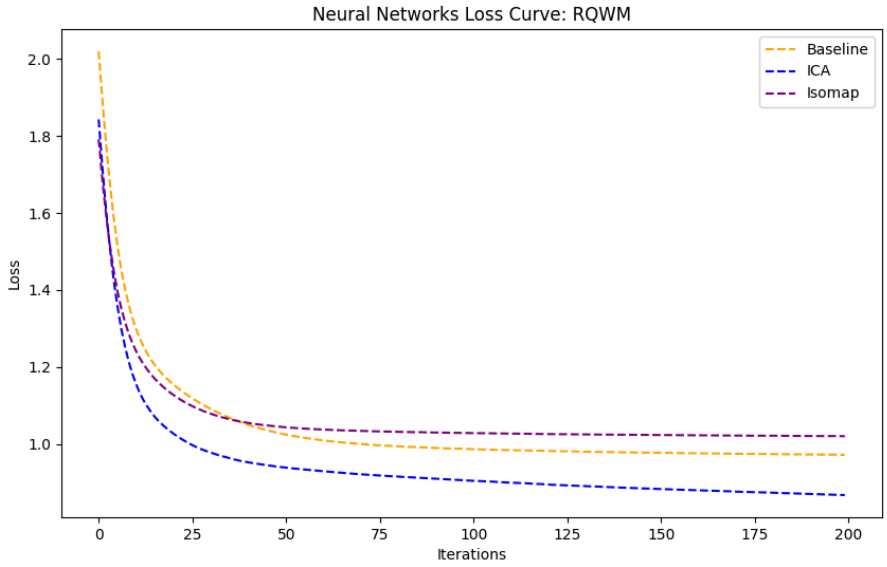

Figure 5—The Loss Curve of Baseline (A1), ICA and Isomap with optimal algorithm components applied.

In terms of performance convergence on the learning curves, the Dimensionality Reduction algorithms ICA and Isomap were around the baseline f1_score, but converged slightly lower. With these values being so close to the Baseline and quantitatively similar, they show high collinearity once again for the RWQM. However, it was surprising and opposite of the hypothesis to note that the ICA faces less loss as it learns data over time than the Baseline.

Table 1—Max Clock Times for Baseline, ICA and Isomap.

	Baseline	**K-Means**	**EM**
Max Fit Time (s):	.332	.315	.275

The clock times were distinctly faster than the baseline and agreed with the hypothesis that a reduction in component makes the algorithm faster

6.2 Clustering Algorithms

The cluster assignments were used as one column added to the current dataset. The values were then normalized to fit the already normalized dataset.

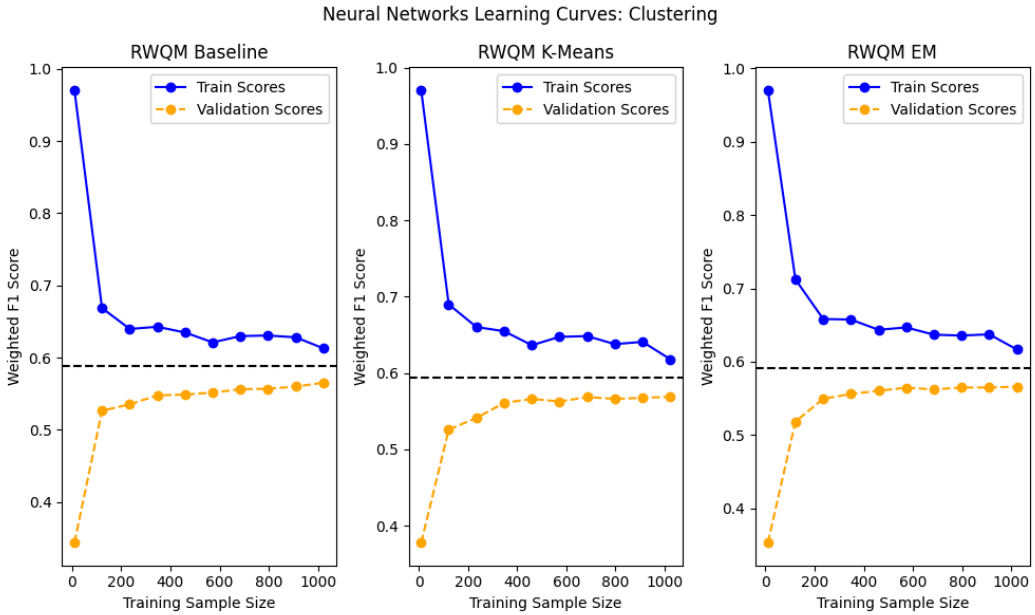

Figure 6—The Baseline (A1), K-Means and EM Learning Curves with optimal algorithm clusters applied.

The hypothesized results came to fruition here with a convergence increase with the learning curves for K-Means and EM, even if it is slightly better performance. It was expected because of the added complexity to the data.

Table 2—Max Clock Times for Baseline, K-Means and EM.

	Baseline	K-Means	EM
Max Fit Time (s):	.264	.263	.263

The clock times were relatively similar, which was not expected. The hypothesized clustering methods should have increased the fit times with added complexity.

7 CONCLUSION

Overall, the dimensionality reductions and clustering algorithms did not overwhelmingly affect the convergence performance of these specific datasets, but they did effectively characterize the data and reduce expensive costs in regards to run time and memory. Each experiment demonstrated how imperative it is to preprocess data and efficiently represent it.

One experiment I would have liked to complete would have been the comparison of the clustering array as the entire training dataset instead of adding it to the current dataset. Especially since it aligned well with the target data, it could have produced increased performance. I do not think the algorithms reached peak cluster or component optimality either because the default algorithms were used without hyperparameter tuning. The tuning of the parameters could have introduced more defined outputs, but the default algorithms were used intentionally to demonstrate how powerful the clustering and dimensionality reduction algorithms can be when appropriately and meticulously exercised on the datasets.

Throughout the tests, all hypotheses held true for these experiments, even when they only helped characterize outputs. From the final results, the algorithm that provided the most information and improved results is EM. The algorithm that provided results, but did not truly vary performance would be PCA. In conclusion, feature selection and feature transformation are enlightening to the preprocessing of continuous variable data.

8 REFERENCES

1. Bohanec,Marko. (1997). Car Evaluation. UCI Machine Learning Repository. https://doi.org/10.24432/C5JP48.
2. Cortez,Paulo, Cerdeira,A., Almeida,F., Matos,T., and Reis,J.. (2009). Wine Quality. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.
3. Kenton, W. (n.d.). *Kurtosis definition, types, and importance*. Investopedia. https://www.investopedia.com/terms/k/kurtosis.asp
4. *2.2. Manifold Learning*. scikit. (n.d.). https://scikit-learn.org/stable/modules/manifold.html
5. *Choose principal components*. Choose Principal Components • SOGA-Py •. (2023, May 5). https://www.geo.fu-berlin.de/en/v/soga-py/Advanced-statistics/Multivariate-Approaches/Principal-Component-Analysis/PCA-the-basics/Choose-Principal-Components/index.html#:~:text=The%20Kaiser's%20rule%20(Kaiser%2DGuttman,retained%2C%20whose%20variances%20exceed%201.