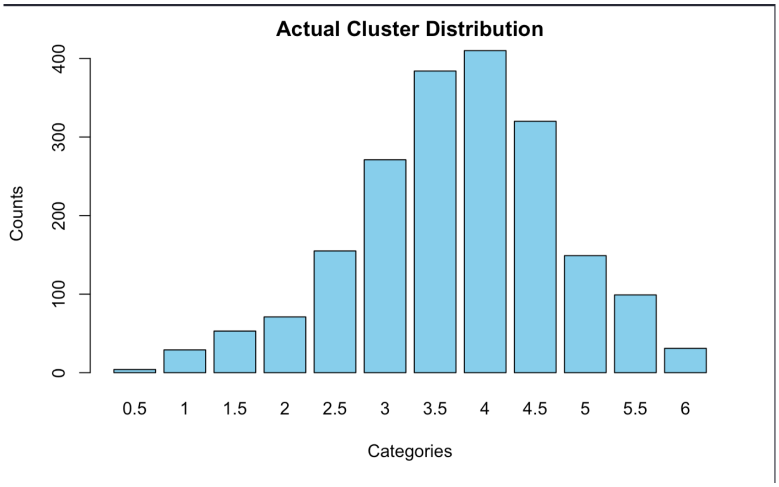
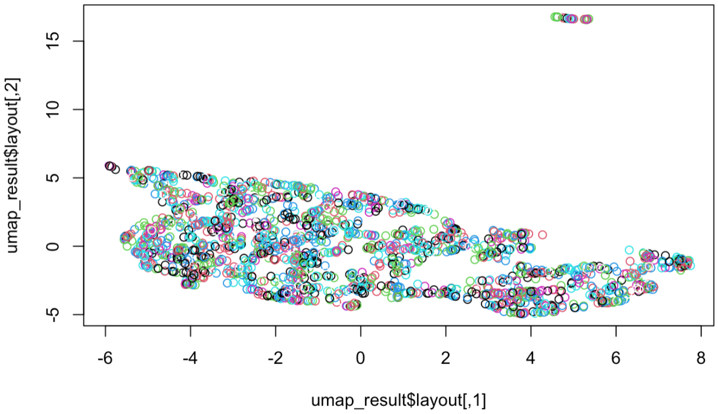
For this data, as we have 17 predictors, and for the response, 12 unique labels: scores from 0.5 to 6 in increments of 12. We can already see a problem of high dimensionality which could affect our model. Before we jump to these conclusions, we can first fit several unsupervised learning models to see if we can glean any insights into the natural clustering of the data. The clustering algorithms we chose to use are the UMAP (Uniform Manifold Approximation and Projection), a K-Means clustering algorithm, and then we apply Principal Component Analysis to the data in order to reduce dimensionality, before fitting a second K-Means clustering algorithm.

We first want to know the actual cluster distribution for reference so we have a sense of how each model performs. I've plotted it on a bar chart to visualize how the scores are distributed.



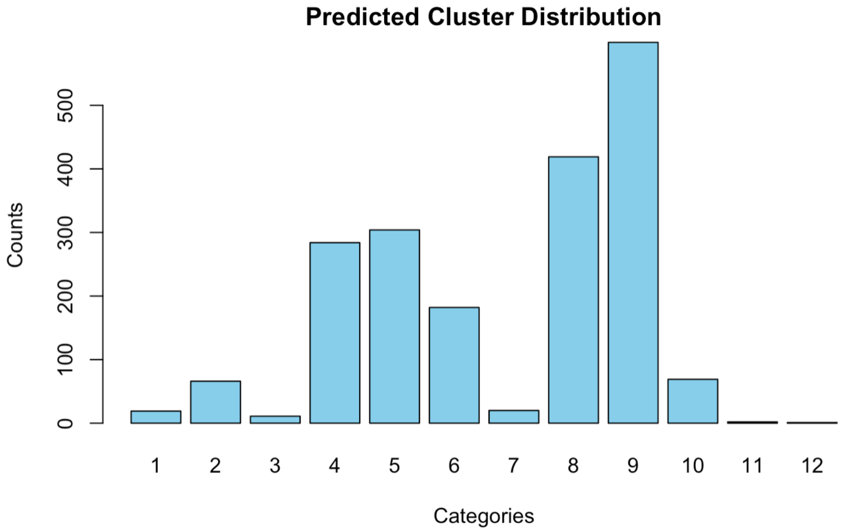
First, we fit a UMAP model to gain insights into the data.



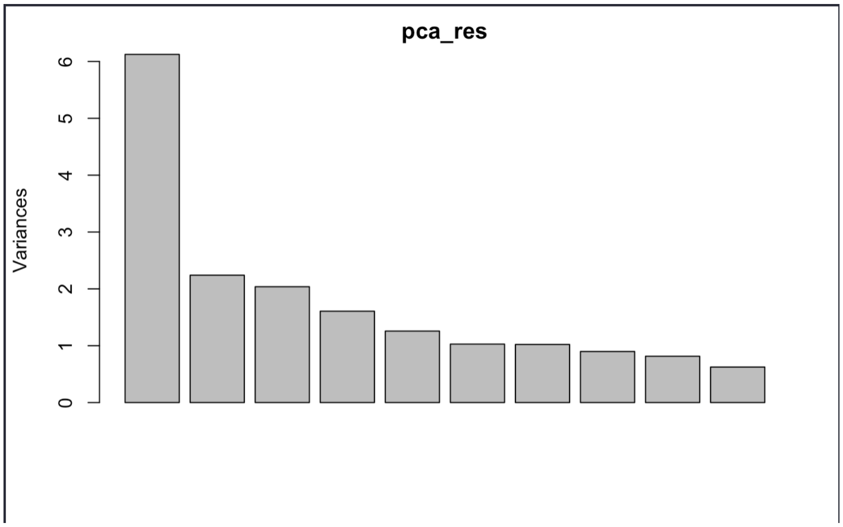
We can see that this data does not form nice clusters by itself, and it may lead to bad results without any form of dimension reduction. We can see this for ourselves by first fitting a preliminary K-Means clustering model with this data. We can try plotting the centroids by cluster label, but due to the high number of predictors, this is not too useful.

As we can see here, there is a very high dimensionality, which may possibly lead to issues with our model. Due to this aforementioned high dimensionality, we can take a look at the distributions of cluster labels to see if this model did a good job. If the distribution is similar, we can correctly assume the model performed adequately. To accomplish this, we use a bar chart that plots the counts of each of the labels.

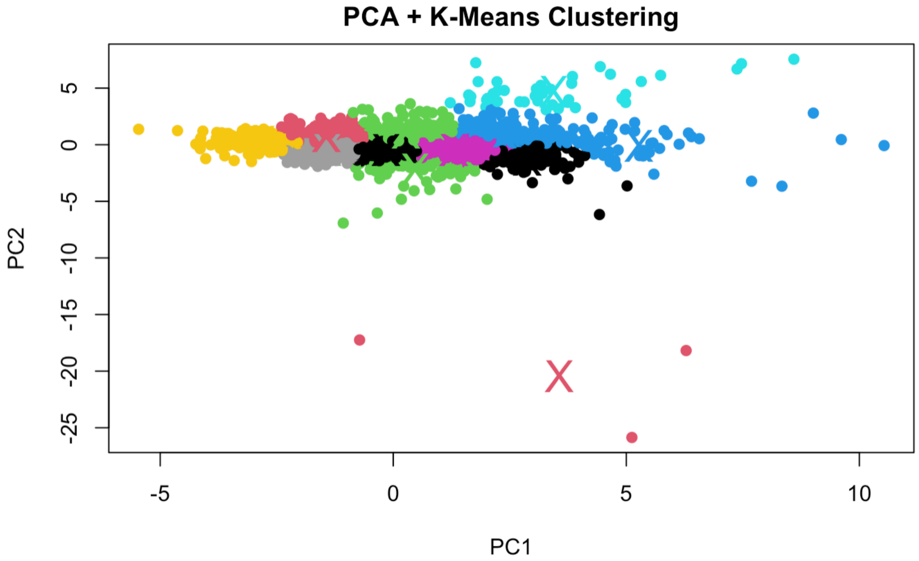
K-Means Clustering

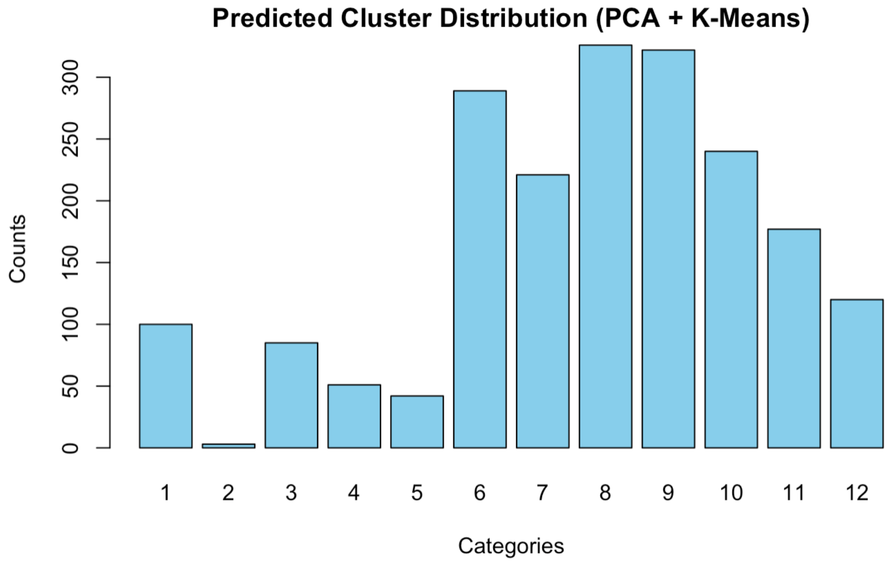


From this bar chart, we can see that the general shape is followed, but the predicted cluster distribution is definitely lacking some information. Therefore, simply the K-Means model may not be enough to adequately represent this data. We saw an issue with dimensionality, as the data has seventeen predictors, so using Principal Component Analysis as a method of dimensionality reduction, followed by a fitted of a new K-Means model may lead to better results. After running the PCA, we can take a look at the principal components ordered by how much variance they explain.



As we can see from the graph of principal components, the ones that explain the most variance in the data are approximately the first 2-4 components. For our purposes, we will retain the first 2 components as our new reduced dataset. We then fit a K-Means model with this data, utilizing 12 clusters as a baseline as that is how many unique scores there are. After fitting a K-Means model with the reduced data, we can produce the clustering plot as follows, and we can also take a look at the predicted cluster distributions.





From this plot, we can see that the reduced data clusters much more nicely, with a few exceptions here and there. Furthermore, most of the centroids were accurately predicted using K-Means, and the data forms well-shaped clusters, which indicates a better model fit. The dimensionality reduction offered by Principal Component Analysis definitely helped, and allowed for a better fit. However, it is clear to see that this data works poorly with clustering algorithms due to the sheer number of predictors and labels, which is why supervised learning may be a better approach to model this data.