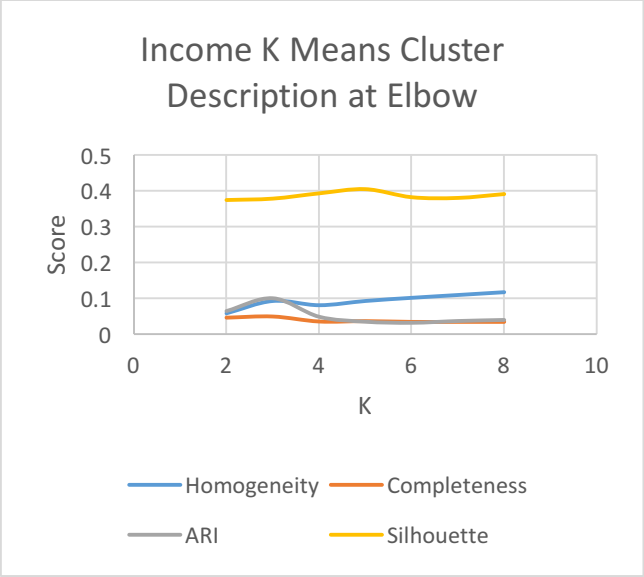
David Vinegar / dvinegar3

An Analysis of Unsupervised Learning Techniques

The two datasets used in this assignment are the same used in assignment 1, the income dataset as well as the housing data set. Income data shows whether or not someone makes more than $50,000 annually, while the housing data shows how much houses cost within Kings County in California.

After running K-means clustering algorithm on the two datasets, the clusters returned are both mostly not homogeneous and mostly not complete. Furthermore, the housing data has a particularly high silhouette score, indicating smaller intercluster distances and larger intracluster distances. On the other hand, the income data has a considerably lower silhouette score at its elbow point (k = 6) than the housing data has at its elbow point (k = 3).

Figure 1



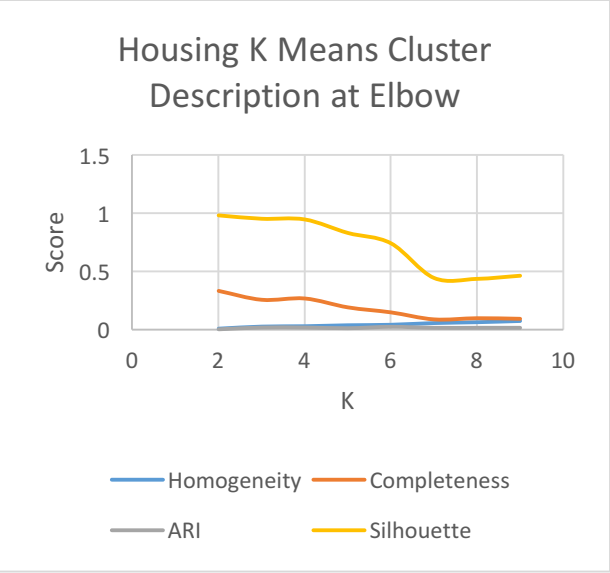
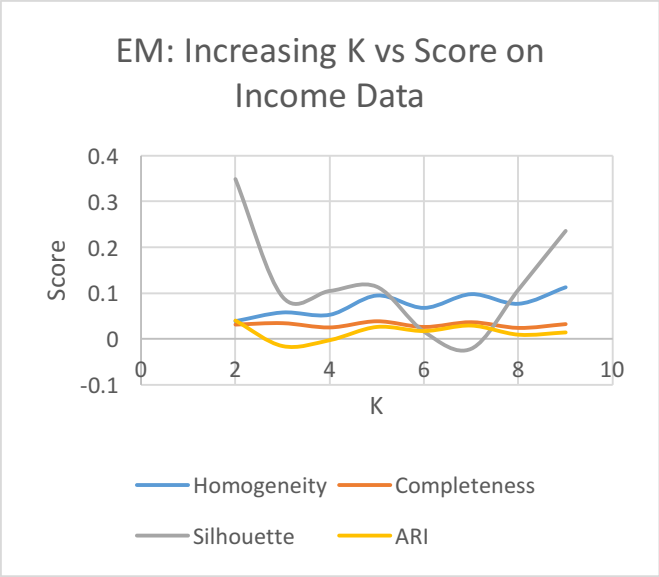
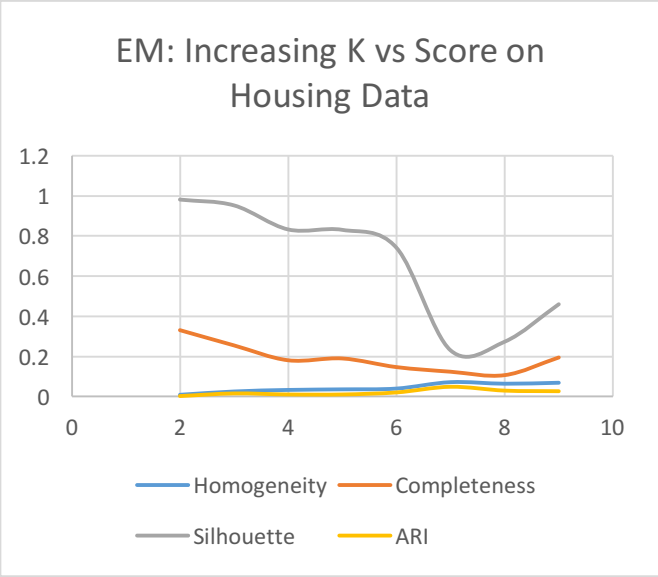


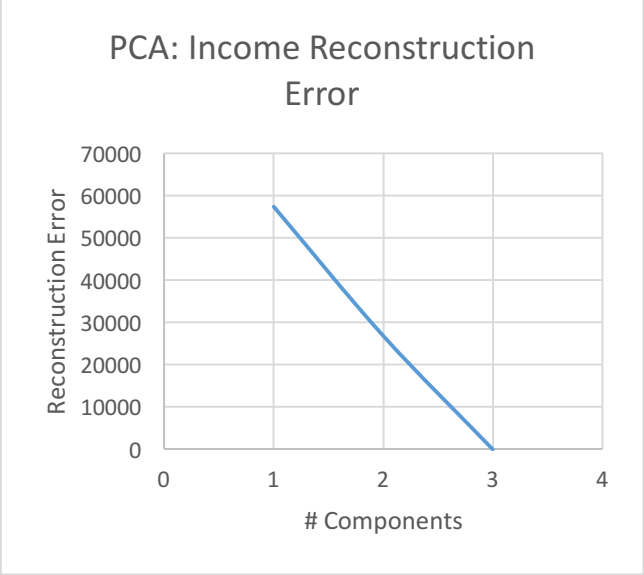
Figure 2

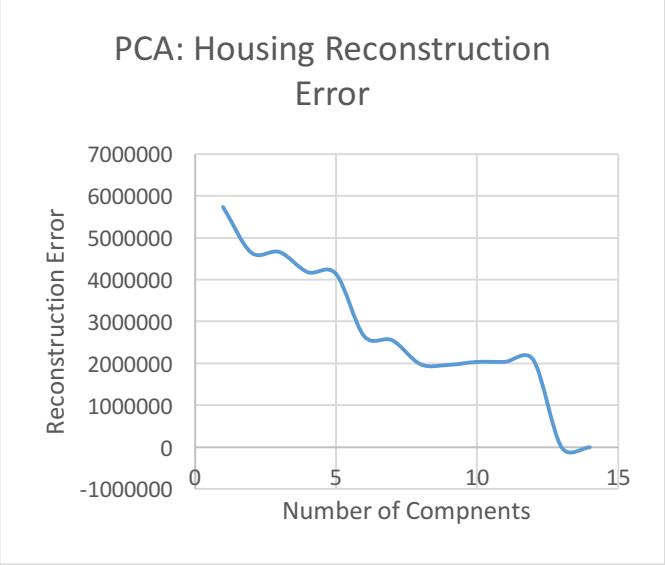
Homogeneity and completeness were slightly higher with the expectation maximization (EM) algorithm, although they were still quite low. Similar to the K-Means clustering algorithm, EM produced a much higher silhouette score on the housing data than the income data.



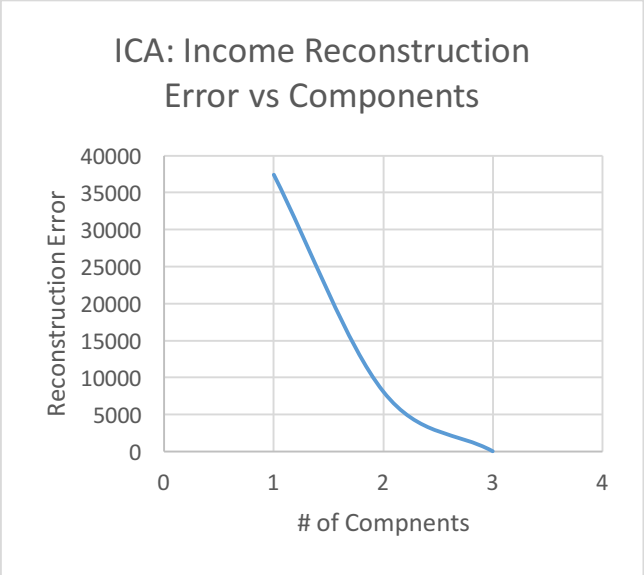


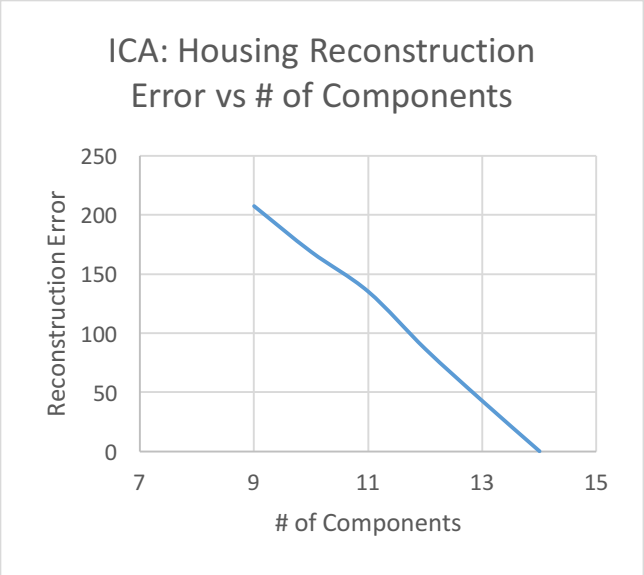
What we are seeing for Principal Components Analysis (PCA) for the income data set is that the data can be transformed linearly into two useful dimensions. The first eigenvalue is accountable for 55.59% of the variance and the second for 42.57% of the variance. The third dimension accounts for significantly less variance, coming in at 1.83%. This distribution of eigenvalues would lend itself to argue for setting the number of components to 2, however, the reconstruction error is much lower with the number of components set to 3 rather than 2. For the housing data set, only one dimension was deemed to be a significant component, accounting for 98.62% of variance. A second eigenvalue came in at a ratio of 1.37%, and the other 12 components rounded down to 0 percent. In both datasets, experimenting with changing the number of components to break down to was conducted, however, when complete, the reconstruction value rose significantly after reducing components. Reconstruction error is lowest when the number of components is equal to 13 for the housing problem and 3 for the income problem.



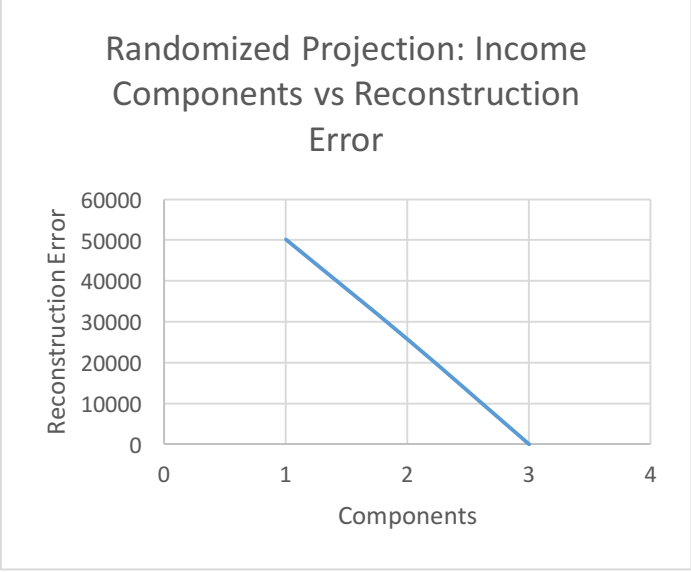


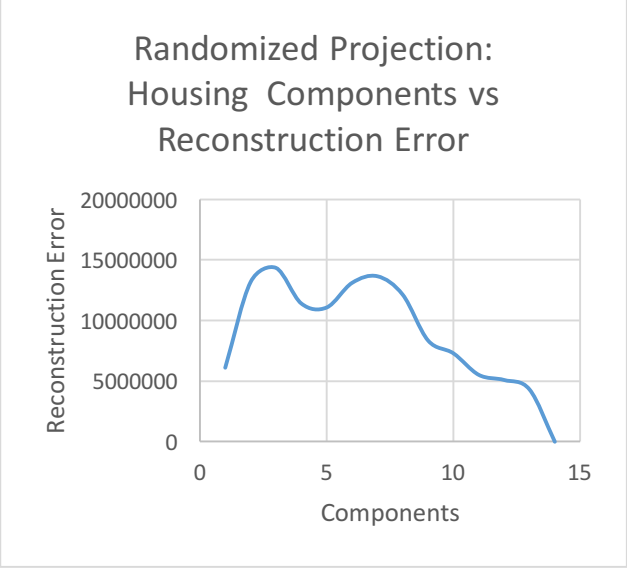
For ICA, the reconstruction error for the income dataset was found to be lowest at 3 components at 2.13E-11, while the housing dataset reconstruction error was considerably higher at 3.12E-06. Still both are very low errors.





Randomized projection led to similar results as ICA and PCA. Its reconstruction error was lowest at k=3 and k=11 components for income and housing, respectively. We can see that the data is reconstructed very well at those optimal levels, with reconstruction error being almost 0. This is very similar performance to what we saw with PCA, where reconstruction error also went down to almost 0.





For feature selection a univariate select best K algorithm was used. The results looked like this:

Income - Ratio of K Best Scores Per Attribute

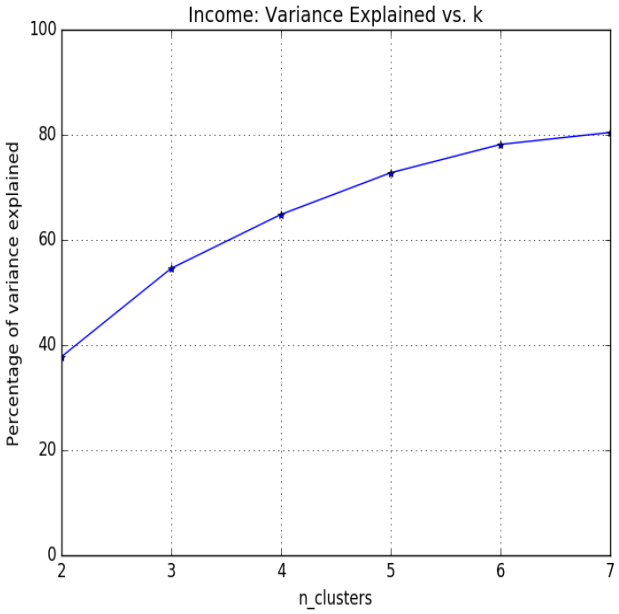
age	0.502320454
education	0.114717652
hours worked	0.382392174

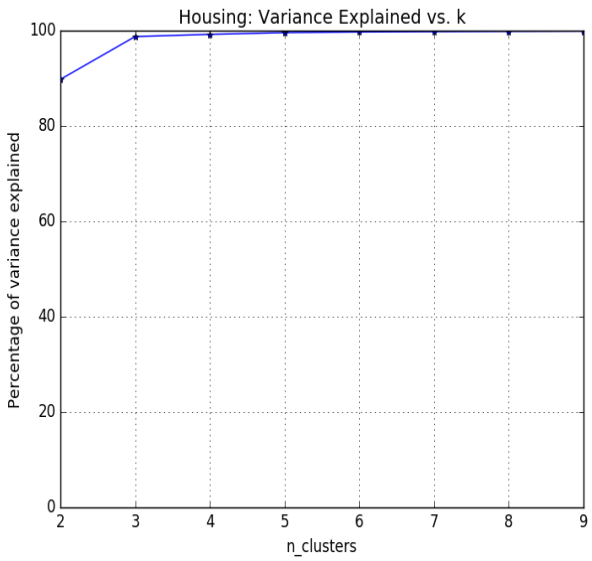
Housing – Ratio of K Best Scores Per Attribute

bedrooms	0.00000	**condition**	0.00000	**yr_renovated**	0.00328
bathrooms	0.00000	**grade**	0.00000	**zipcode**	0.00000
sqft_living	0.00240	**sqft_above**	0.00196	**sqft_living15**	0.00196
sqft_lot	0.83980	**sqft_basement**	0.00157	**sqft_lot15**	0.14901
floors	0.00000	**yr_built**	0.00000		

We can see that feature selection found two particularly useful attributes for the housing dataset, namely sqft_lot, which denotes how many square feet the lot that the house sits on is, and sqft_lot15, which denotes the average square footage of the fifteen closest lots nearby. In other words, what matters most is how big the house (and its surrounding land) is and how big the neighbors' houses and surrounding land are. For the income dataset, all three attributes were found to be important, but age and hours worked were considerably more important than education, which is quite surprising. It would seem more likely that education would play a greater role.

K was chosen using the elbow method. For both datasets, k was chosen to be wherever the elbow was. For the income dataset, k was found to be best at 6, while for the housing dataset, k was found to be best at 3. I also examined multiple metrics concerning the clusters created, such as homogeneity, completeness, and the silhouette score, but found the best results to be at the elbows, which was particularly useful for EM as well. Optimal number of components for dimensionality reduction algorithms was chosen using the best reconstruction error.





Analyses of results

The clusters, by most descriptory measures for both problems and clustering methods, were unimpressive. The income dataset, as described in figure 1 did poorly on each descriptory measurement. Homogeneity for KM came in at 0.093, completeness at 0.049, silhouette at .378, and inertia at 6.67E5 versus EM with .094 homogeneity, .038 completeness, and .114 silhouette score. It could be expected that income data would cluster well, that similar income, education, and hours worked levels would end up clustered together. One possible explanation for the poor results could be that instances that are similar within one attribute, such as education, were often dissimilar in other attributes, such as hours worked. Additionally, it is known from my experiments on this dataset for supervised learning that getting an accuracy rate over 80% is very difficult. Unsupervised learning knows less about the problem than supervised learning, so putting 80% of examples into the correct class serves as a good baseline. It makes sense that clustering would be worse than any supervised learning technique. The completeness score of 0.049 for KM and .038 for EM lends evidence to the hypothesis that clustering should perform worse than supervised learning techniques because this means that far fewer than 80% of either class ended up in a single cluster.

Similar to the income dataset, clustering did not work particularly well on the housing dataset. Homogeneity, completeness, and inertia for KM scored quite poorly at 0.025, 0.254, and 3.94E10, respectively, while homogeneity and completeness for EM were also bad at .25 and .254 . Of note is that the silhouette score for EM housing was similar to the silhouette score to KM for housing. The same pattern was observed in the income set; its silhouette score was similar for KM and EM. Perhaps this happened because the clustering

abilities of both algorithms are similar. Like what we observed previously with the income dataset, the housing dataset results were likely received because instances similar in certain attributes were probably very dissimilar in other attributes. With labeled data and supervised learning, results accuracy around 67% was possible. The clusters, as described by my metrics, do not line up with the supervised labels. With a completeness score of 0.254, it is clear that most classes don't end up in the same cluster, and hence it is not possible that the clusters could be all lining up naturally, or even mostly lining up naturally with the class labels.

Performance differences due to the problem at hand were existent, but definitely not as pronounced as they were with supervised learning. For starters, clustering techniques performed poorly on both problems, whereas supervised learning techqniues performed well on both problems. On the other hand, Homogeneity, completeness, and ARI were very comparable for both problems for EM and K-means. Inertia for K-means clustering was significanlty higher for the housing problem than the income problem, potentially due to the significantly larger number of features. As shown by the reconstruction error graphs earlier, results were vastly different between the two problems with PCA, ICA, and RP; the reconstruction error is on different orders of magnitude. Feature selection found all three features to be relevant to at least 10% for the income dataset, whereas the sweeping majority of features were found to account for less than .0000% of the variance in the housing dataset.

K-means clustering was able to be improved by adjusting the number of clusters. It was found to be approximately optimal at around n = 6. This was not exact, because measures of optimality can conflict. For example, homogeneity improves as the number of clusters increases, while completeness decreases. Other parameters for k-means were found to largely not make a difference, such as number of iterations, the number of times the k-means algorithm runs with different centroid seeds (n_init), and the method for initialization. It makes sense that the number of clusters would be the most important factor, since they are directly related to multiple descriptory statistics such as homogeneity and the silhouette score.

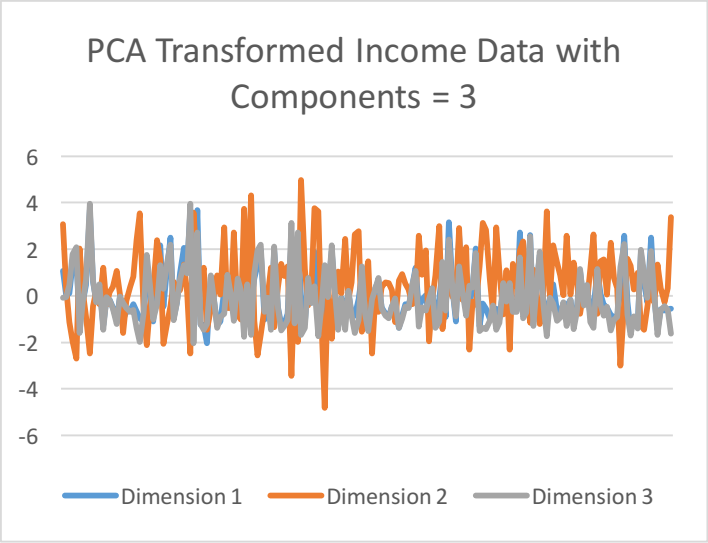
Like KM, EM was also found to be largely insensitive to most parameters for both problems. The only parameter that substantially changed the asnwer was the number of components. The covariance type, the convergence threshold, and the number of iterations had no effect on the results received.

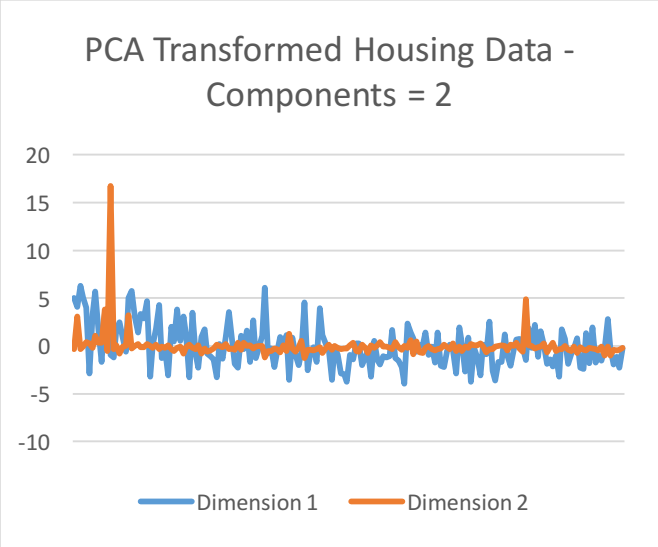
Both PCA and ICA were very sensitive to the number of components. Within PCA, parameters such as tolerance and whiten proved ineffective at changing the results. The solver used in PCA was able to very marginally improve the outcome when set to 'arpack'. For ICA, changing its algorithm from parallel to deflation failed to improve results. Changing tolerance and iterations were also ineffective.

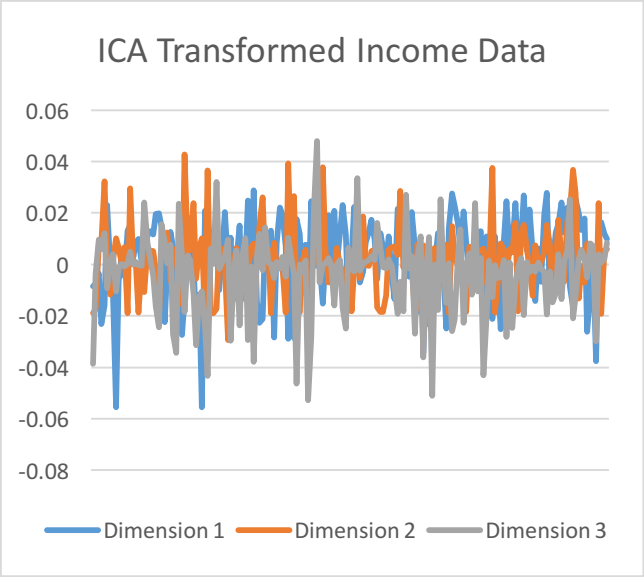
RP had only a couple parameters available to change, but similar to the other algorithms, the vast majority were largely ineffective at improving or even changing the results. Eps, or the parameter to control the quality of the embedding, proved not to make a difference in reconstruction error. Changing the random_state parameter, which controls the random number generator used to generate the matrix at fit time, changed the reconstruction error, but not meaningfully.

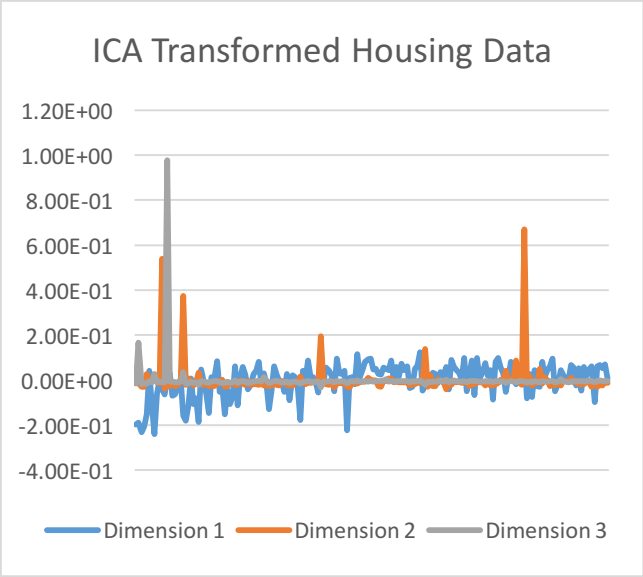
Feature selection using SKLearn's SelectKBest only has one parmater available to change, which is K. Changing K leads to fewer features being selected, which did not alter results.

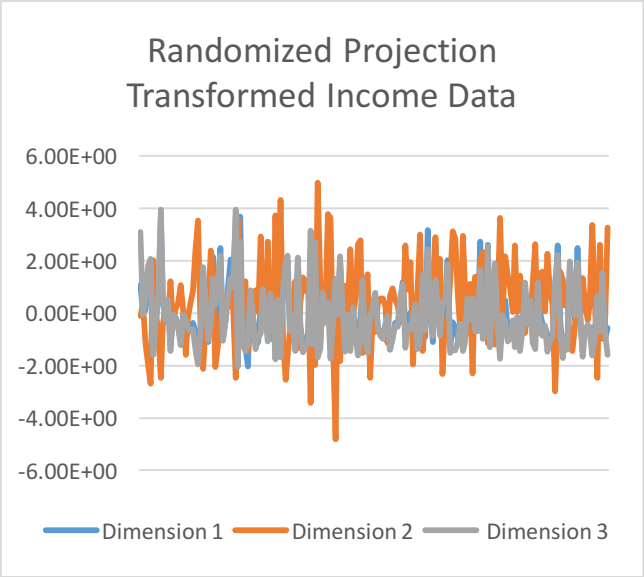
The best way to describe how the data looks in the new spaces created with the various algorithms is to show the trasnformed data in line charts below.

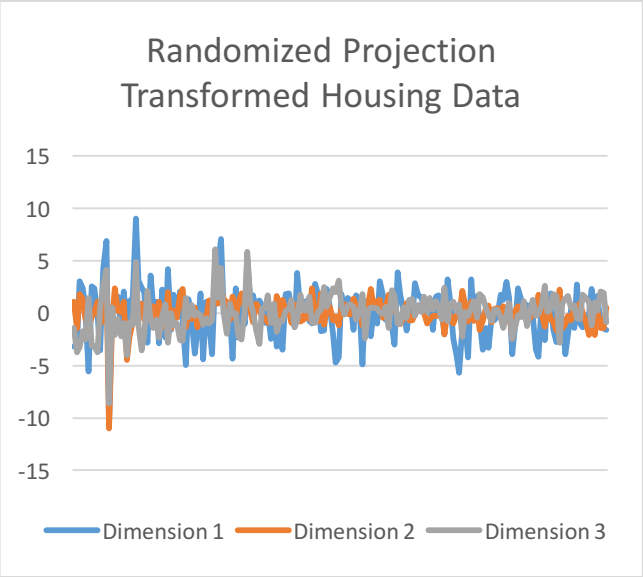






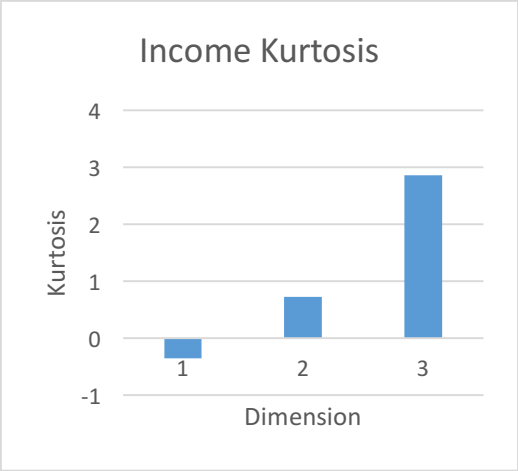


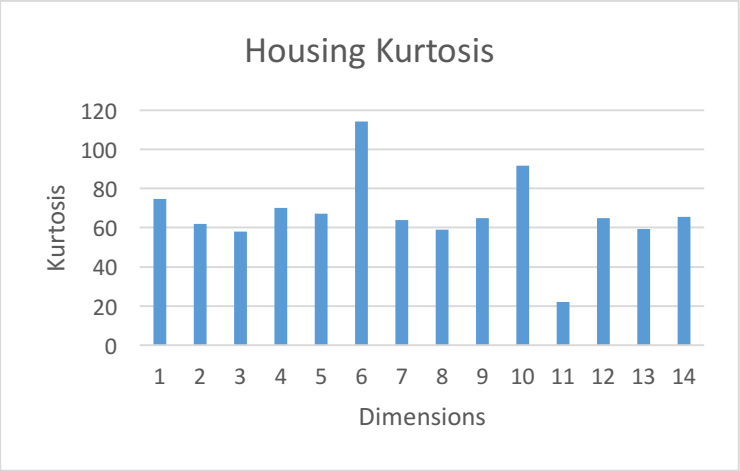




What we see with the income data, across the three algorithms above, is less compressed data, especially in comparison to the housing data. This could be signalling higher variance, which will be revisited when describing kurtosis. The housing data is more compressed, but it seems to have some sharp outliers.

Housing kurtosis was shown to be signicantly larger than income kurtosis after running ICA. We can imply from this that the income data is considerably less fat-tailed, in other words, there are fewer outliers. This is in line with what was suggested previously when examining the data in the new spaces created, where we consistently saw some sharp outliers within the housing data.





The projection axes show much higher variance for the income data set than with the housing dataset. This could add evidence to what was shown previously in the graphs depicting ICA transformed data, that there is more noise in the income dataset.

Variance of reconstruction error was very small when rerunning RP, coming in at 2.35E-18 and 2.25E-14 for income and housing, respectively. Part of the reason for this is that the reconstruction errors, when using an optimal number of components (3 for income and 14 for housing), were exceedingly small. Naturally, variance between a set of small numbers will be small as well.

When reproducing clustering experiments on the datsets projected onto new spaces created by ICA, PCA, and RP, the clusters were largely dissimilar. The clusters were the same with PCA for K-means clustering on both housing and income, but different for ICA and RP. I believe the clusters were the same for PCA, since the number of components used was the same as the "elbow" indicated I should use from doing K-means clustering. ICA and RP don't have that same relationship going on, which means the results should be different.

Income Clustering after Dimension Reduction vs Without

Clustering	Dimension Reduction	Inertia	Homogeneity	Completeness	ARI	Silhouette
K-means	None	321412	0.102	0.035	0.032	0.382
K-means	PCA	320701	0.102	0.035	0.032	0.383
K-means	ICA	1	0.11	0.056	0.085	0.92
K-means	Random	324	0.108	0.086	0	0.285
EM	None	N/a	0.067	0.026	0.017	0.016
EM	PCA	N/a	0.038	0.03	0.04	0.35
EM	ICA	N/a	0.045	0.036	0.019	0.331

EM	Random	N/a	0.126	0.098	0.027	0.212

Housing Clustering after Dimension Reduction vs Without

Clust-ering	Dim Reduce	Inertia	Homog-eneity	Comple-teness	ARI	Silho-uette
KM	None	39.4E9	0.025	0.254	0.017	0.953
KM	PCA	39.4E9	0.025	0.254	0.017	0.953
KM	ICA	25	0.017	0.209	0.011	0.772
KM	Random	62013	0.113	0.091	0.009	0.293
EM	None	N/a	0.025	0.254	0.017	0.953
EM	PCA	N/a	0.116	0.122	0.025	0.315
EM	ICA	N/a	0.118	0.158	0.043	0.02
EM	Random	N/a	0.113	0.092	0.017	0.267

Training the neural network with any dimension reduction algorithm resulted in the same test error rate of 0.24144. This same test error rate was also see when training the neural net with randomized optimization and backpropagation. Clustering usually came in at marginally worse than the other training methods. What is more interesting is the training times. Training time was lowest when training with EM clustered data, coming in at 0.0087s, about half the time required by backpropagation, 0.015s. Dimensionality reduction took substantially longer than backpropagation. This is due to dimensionality reduction being considerably more computationally expensive than clustering and backpropagation.