**Unsupervised Learning**

**Introduction**

This assignment is about exploring unsupervised learning algorithms, containing two clustering algorithms:

* K-means
* Expectation Maximization

Also including four dimensionality reduction algorithms :

* PCA
* ICA
* Randomized Projection
* Factor Analysis

The dataset used for the assignment are:

* Adult Data Set from UCI
* Mushroom Data Set from UCI

I have selected the problem from UCI by filtering on classification problems. These datasets

(Adult Data Set and Mushroom Data Set) are interesting and maybe more straightforward

because they are used to predict only one class.

**Adult Dataset:**

The Adult Data Set is an extract of the 1994 US Census. Using this data you can predict

whether someone’s income will exceeds $50K/yr based on census data. Also known as

"Census Income" dataset. The data set has 48,842 instances and contained some missing

values that were removed. There are 14 attributes and 1 class (income above or below 50k.)

The data set is interesting because we can see which features are relevant to predict

someone’s income. You can easily have a meaningful conversation why a specific factor may

contribute to an income of >50k or <50K.

**Clustering Algorithms:**

Clustering is about grouping a set of objects such that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

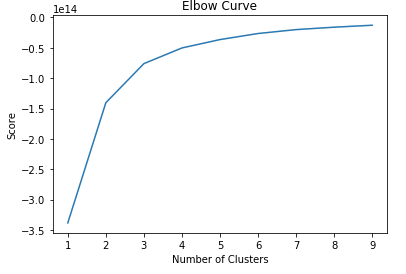
**K-means clustering**

K-means clustering is used with unlabeled data where the algorithm is used for finding groups of data which are similar in some way, where k is chosen.

In the example I am providing, I have chosen the K value based on the elbow curve representation which is gave the best estimate for k as 2.

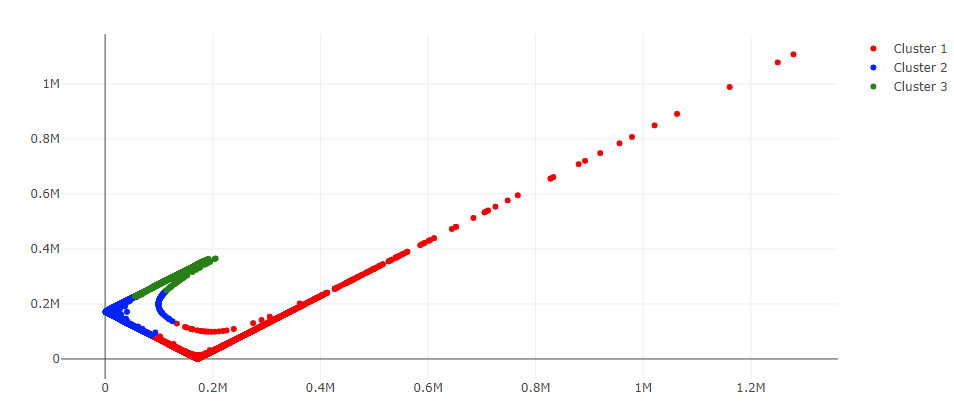
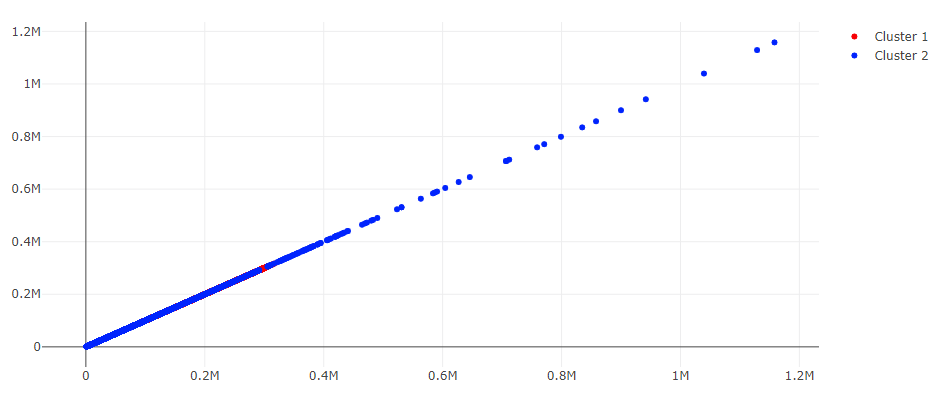
Clusters:2

Based on the elbow curve below we can see that the ideal number of clusters for the adult dataset is between 2 and 4. Higher than the earlier mentioned number of clusters would not add variance to the curve.



Looking at the two clusters below we can assume that the graph with 3 clusters is actually a rotated version of the graph with the cluster 2.

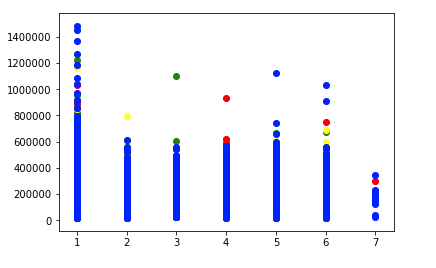
cluster= 2 cluster = 3



**Expectation Maximization**

The Expectation Maximization algorithm is used for finding the maximum likelihood estimates of parameters in probabilistic models which is an iterative method, alternating between two steps, expectation (*E*) and maximization (*M*). In my example for clustering with EM, I have applied the Gaussian mixtures algorithm with 4 components

Components: 4



Converged log-likelihood value: 30.571303321321526

Number of iterations needed until it is converging: 15

**Dimensionality Reduction :** Dimensionality reduction algorithms are used when the number of features are high and it may be useful to reduce it before applying a supervised learning step to the data.

.

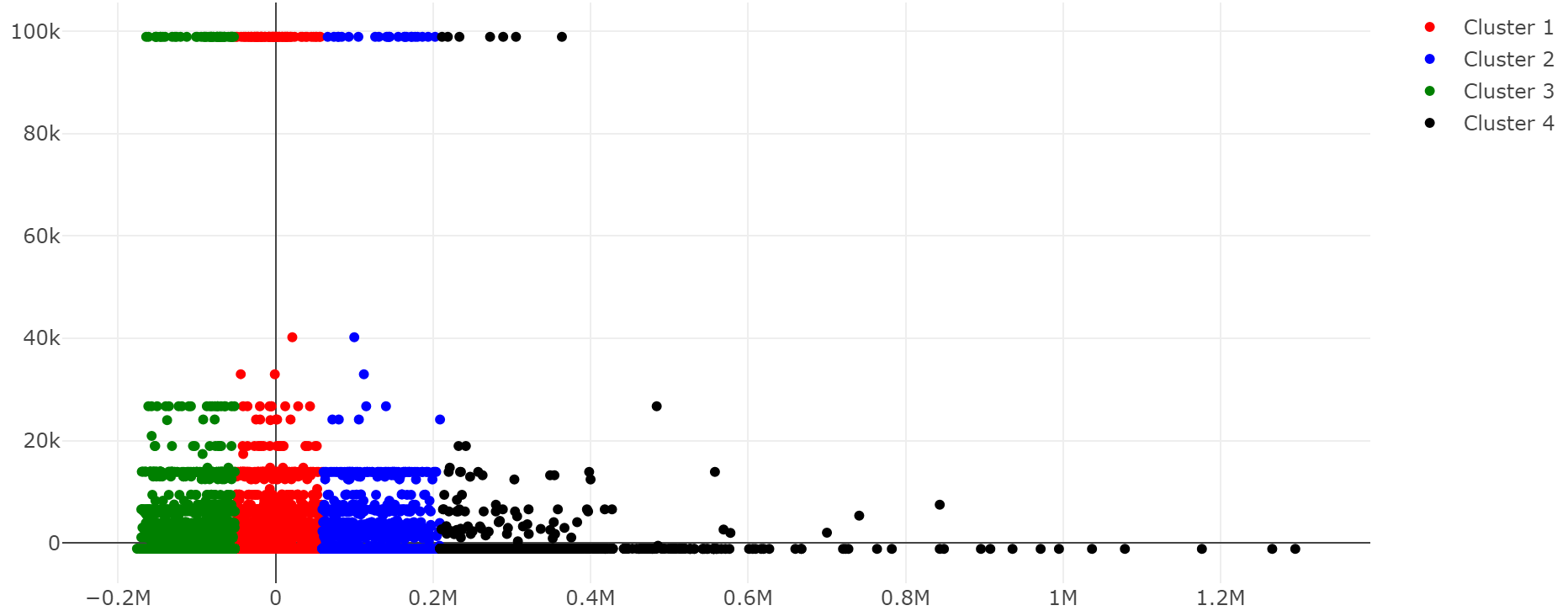
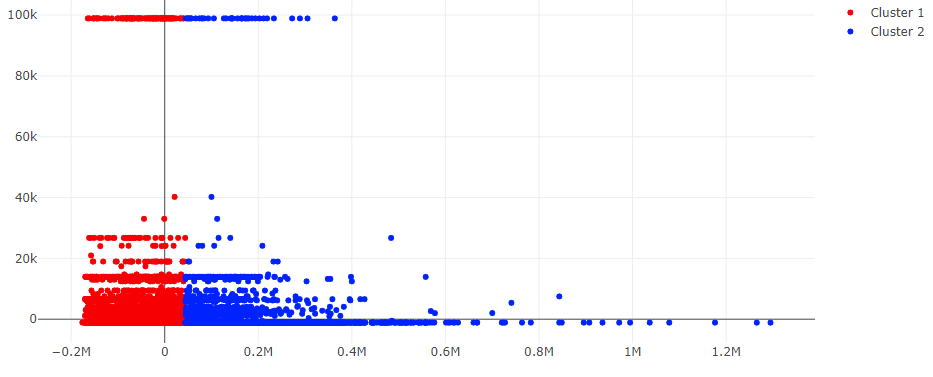
**PCA**

Principal component analysis is a statistical procedure that uses [orthogonal transformation](https://en.wikipedia.org/wiki/Orthogonal_transformation) to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of [linearly uncorrelated](https://en.wikipedia.org/wiki/Correlation_and_dependence) variables called principal components.(Wikipedia, 2019)

In the below graph adding more clusters did not change the shape of the distribution however the data is evenly clustered around centroids.

Components: 3

Clusters=2 Cluster= 4



**ICA**

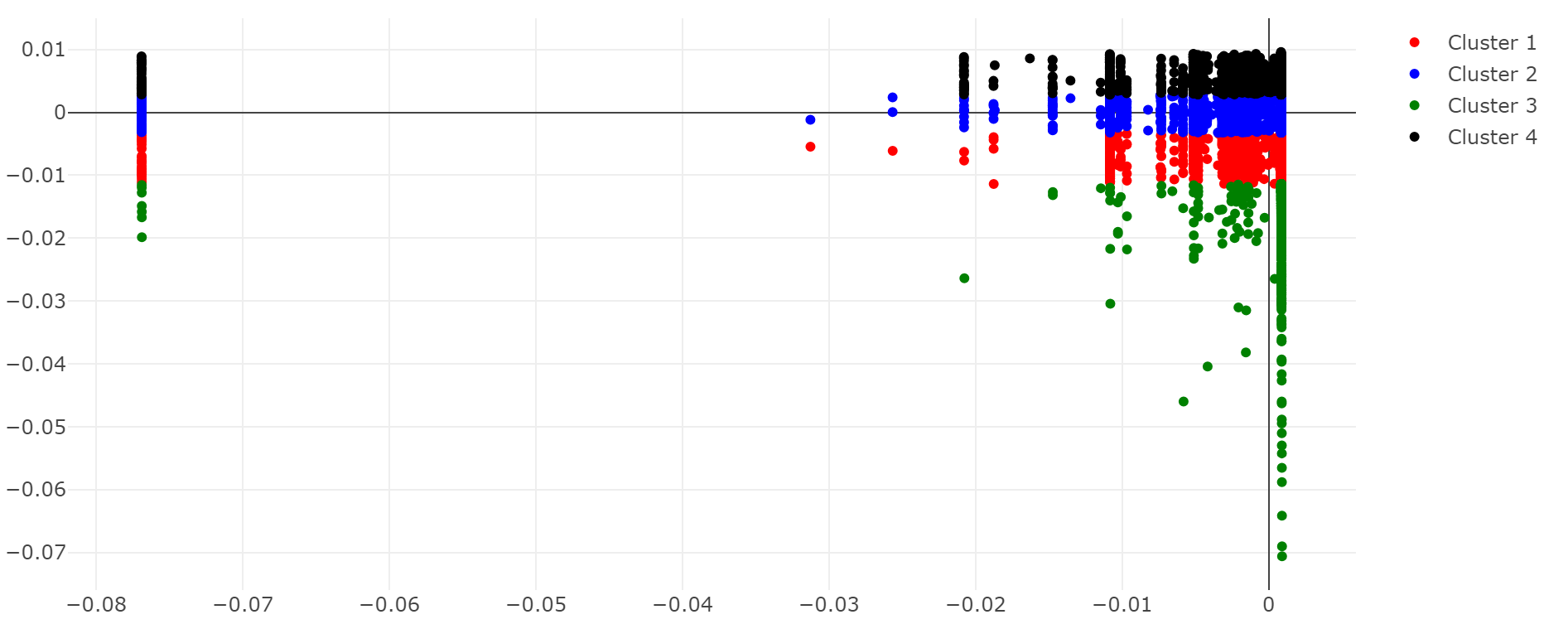
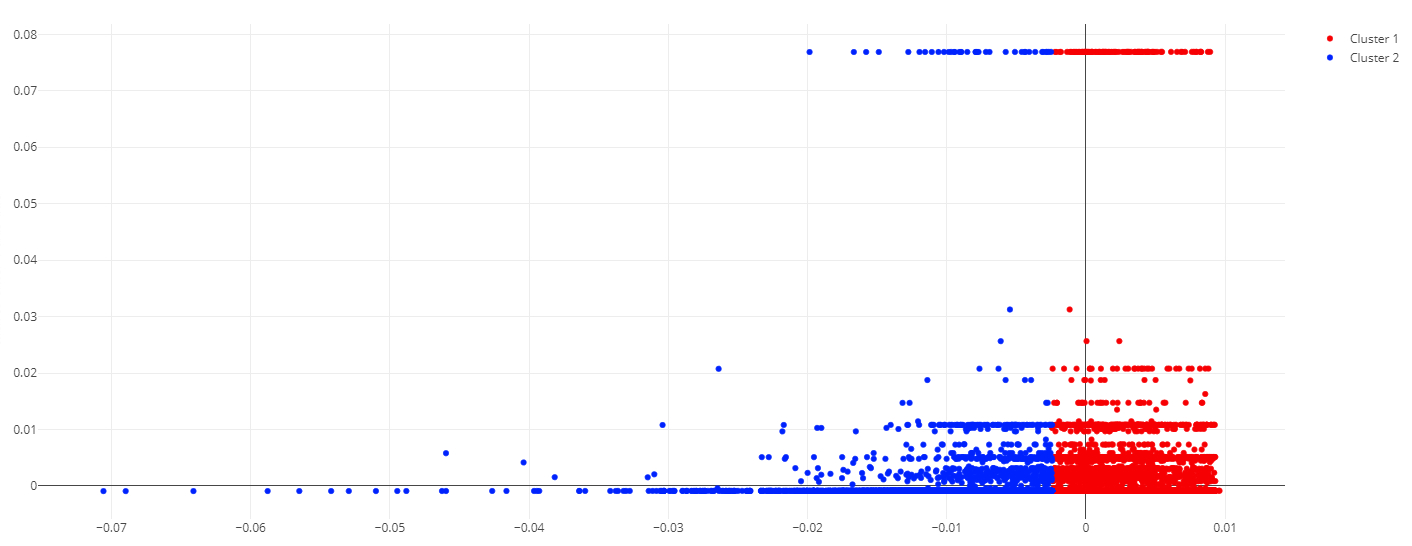
Independent component analysis is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals.

ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples. In the model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed non-gaussian and mutually independent, and they are called the independent components of the observed data. These independent components, also called sources or factors, can be found by ICA.(Hyvärinen, 2019)

In the below graph we can see that the data points are rotated based on the number of clusters defined.

Component: 2

Clusters=2 Cluster= 4



**Randomized Projection**

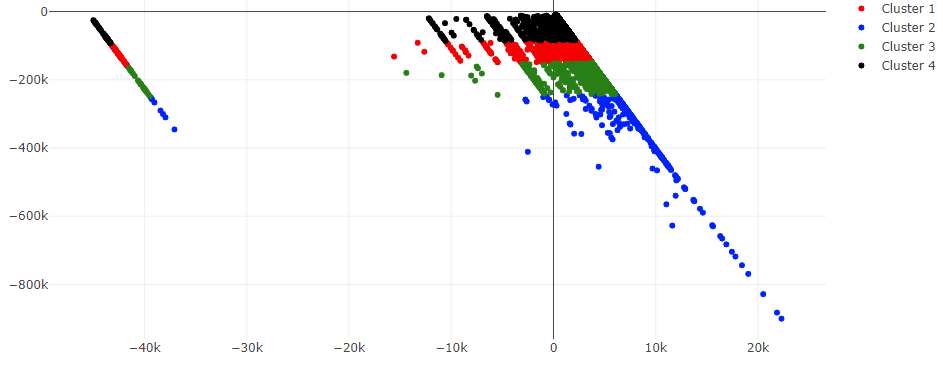
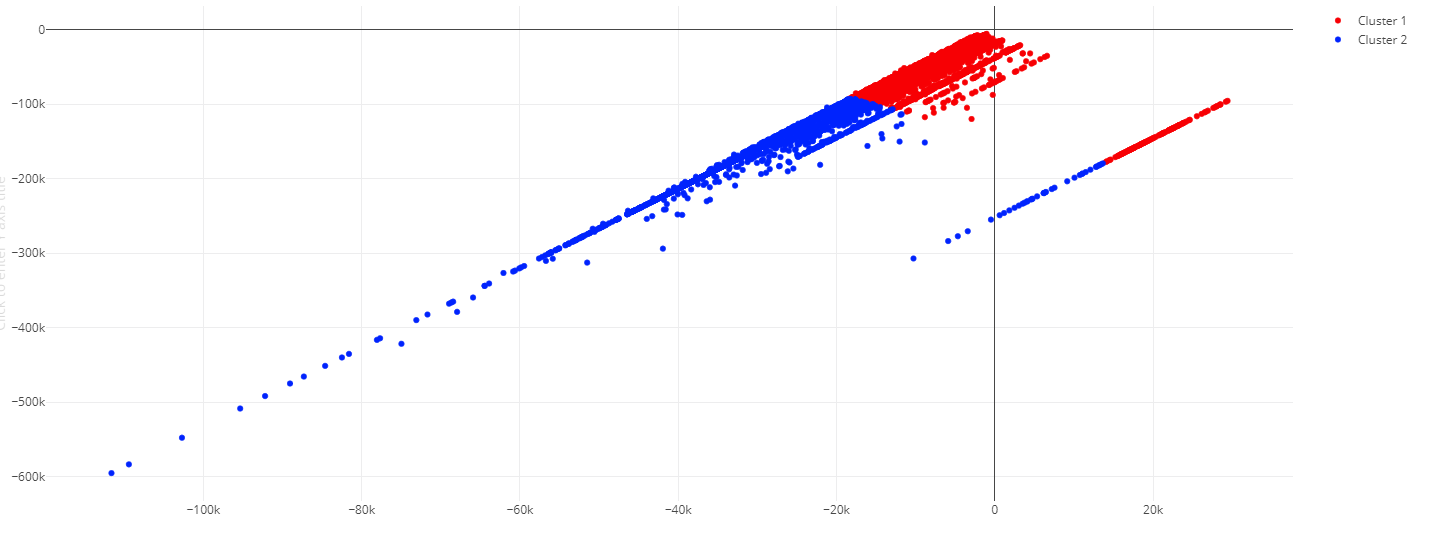
Random projection is used to reduce the dimensionality of the data by trading a controlled amount of accuracy (as additional variance) for faster processing times and smaller model sizes.

The dimensions and distribution of random projections matrices are controlled so as to preserve the pairwise distances between any two samples of the dataset.(Scikit-learn, 2019)

In the below graph we can see that the data points are rotated based on the number of clusters defined.

Components: 5

Clusters=2 Cluster = 4



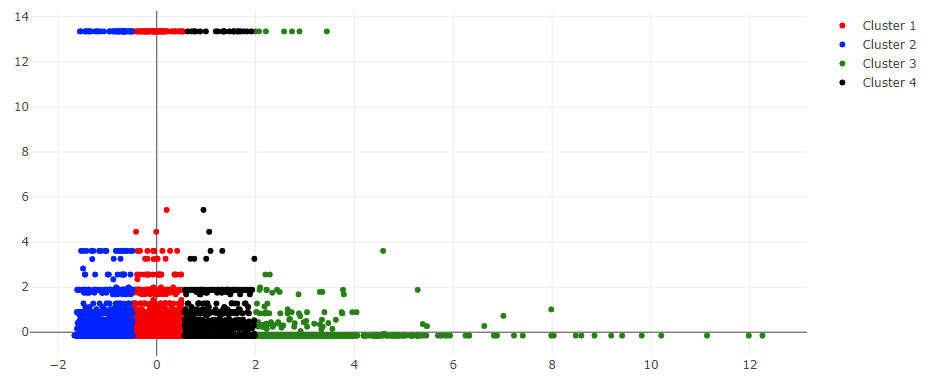
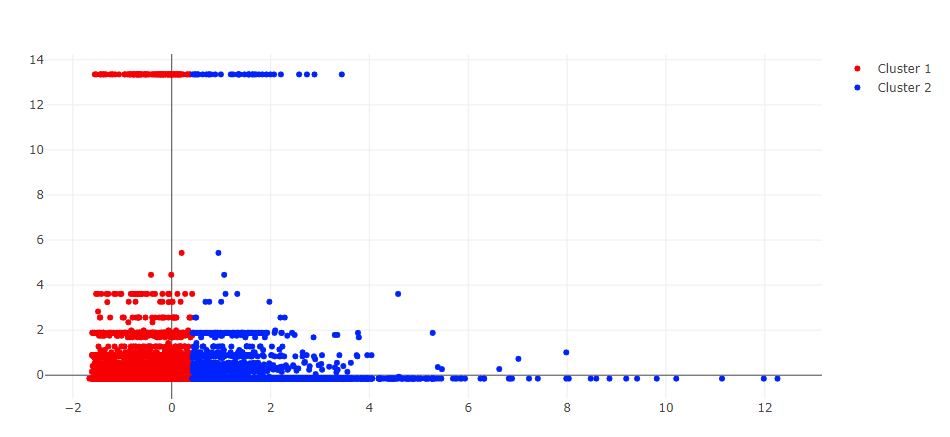
**Factor Analysis**

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. In other words, it is possible, for example, that variations in three or four observed variables mainly reflect the variations in fewer unobserved variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modelled as linear combinations of the potential factors, plus "error" terms. (Wikipedia, 2019)

In the below graph adding more clusters did not change the shape of the distribution however the data is evenly clustered around centroids.

Components: 5

Clusters:2 Clusters = 4



**Mushroom Dataset:**

The Mushroom Data Set describes in terms of physical characteristics if a mushroom is

poisonous or edible. The data set has 22 attributes or features and 1 class (poisonous or not).

Using this data set we can predict based on the appearance of the mushroom if a human can

consume it. The data set contained 8124 instances and I have dropped 1 column (stalk\_root)

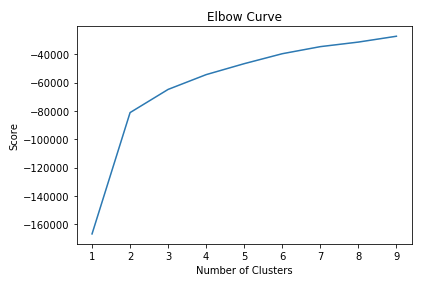
that contains too many null values(2560 null values out of 8124).

**Clustering Algorithms:**

**K-means clustering**

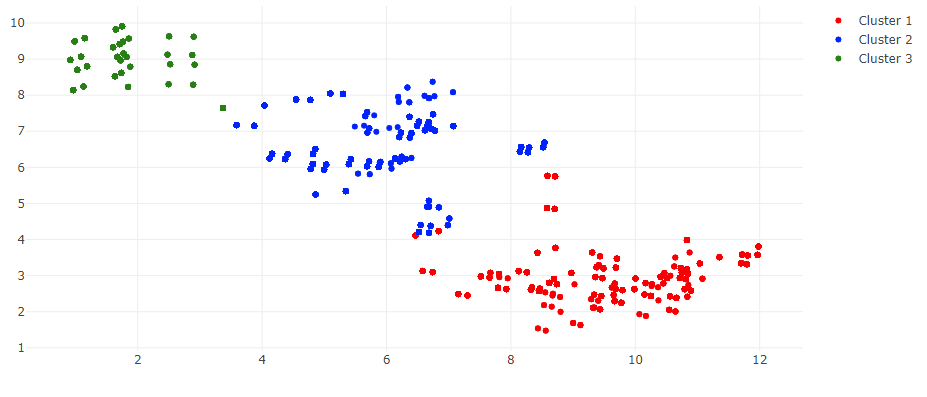
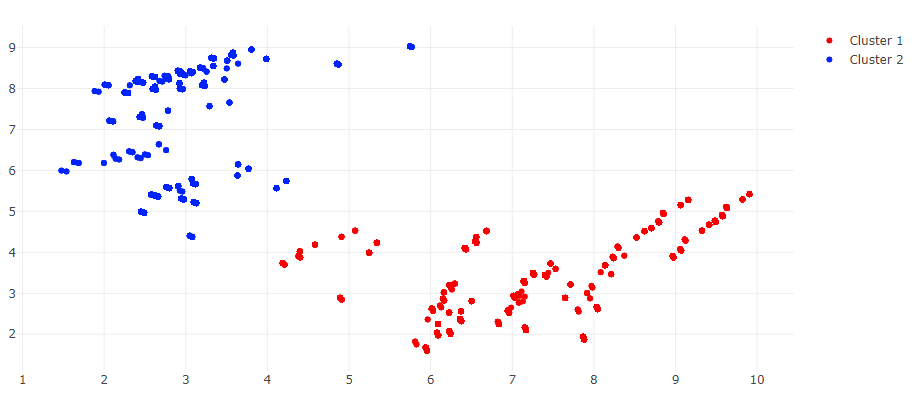
Clusters: 2

Based on the elbow curve below we can see that the ideal number of clusters for the adult dataset is between 2 and 7. Higher than the earlier mentioned number of clusters would not add variance to the curve.

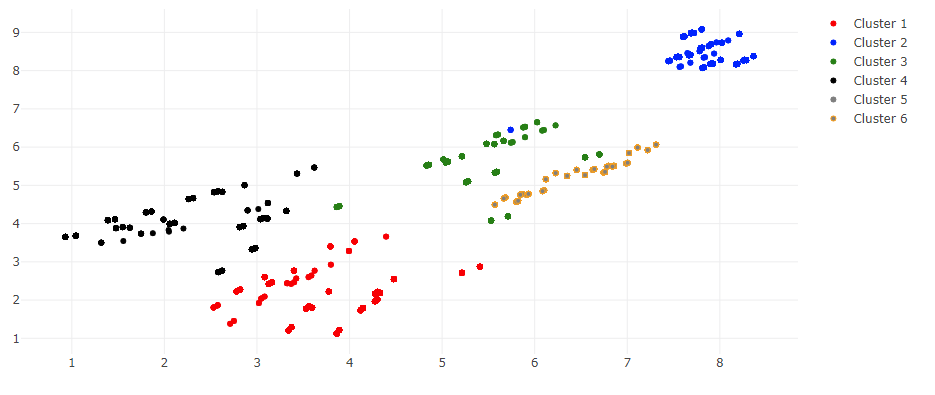


In the below graph adding more clusters change the shape of the distribution and the data is evenly clustered around centroids.

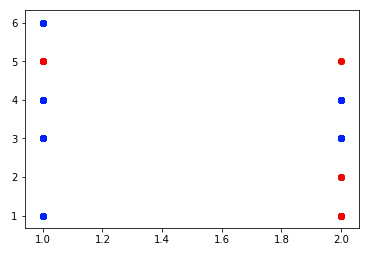
Cluster = 2 cluster =3



Cluster = 6



**Expectation Maximization**



Converged log-likelihood value: 4.639030563439128

Number of iterations needed til it is converging: 7

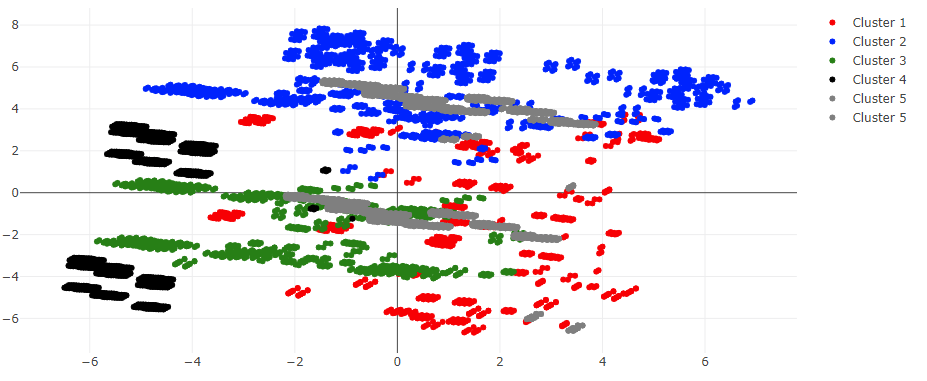
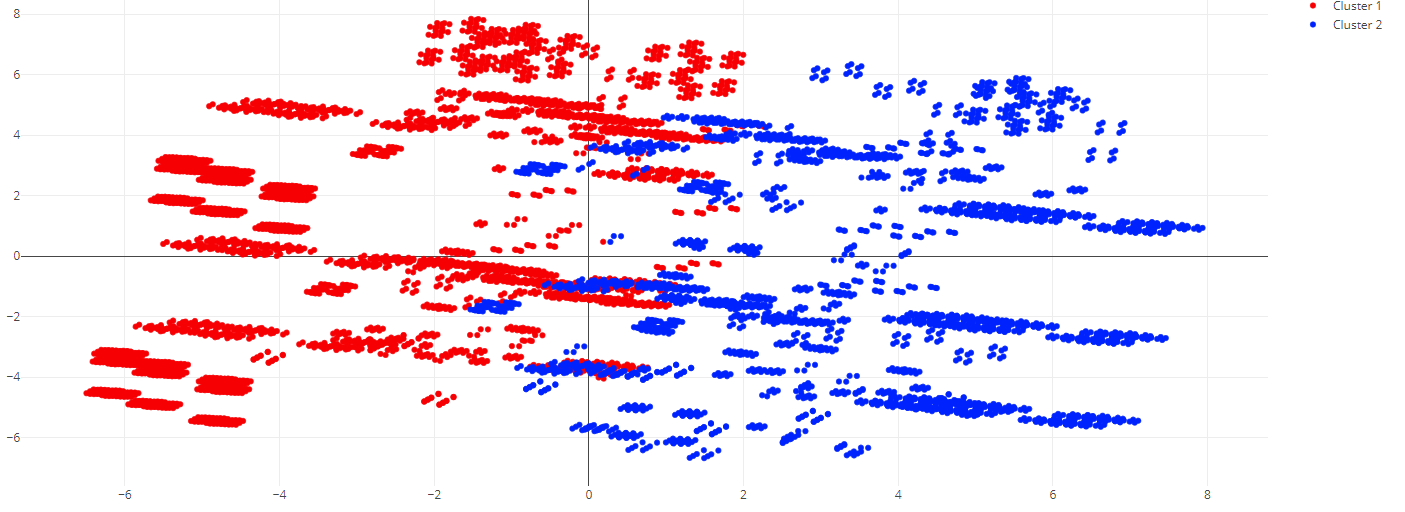
**Dimensionality Reduction :**

**PCA**

In the below graph adding more clusters did not change the shape of the distribution however the data is evenly clustered around centroids.

Components: 2

Clusters:2 Clusters = 6

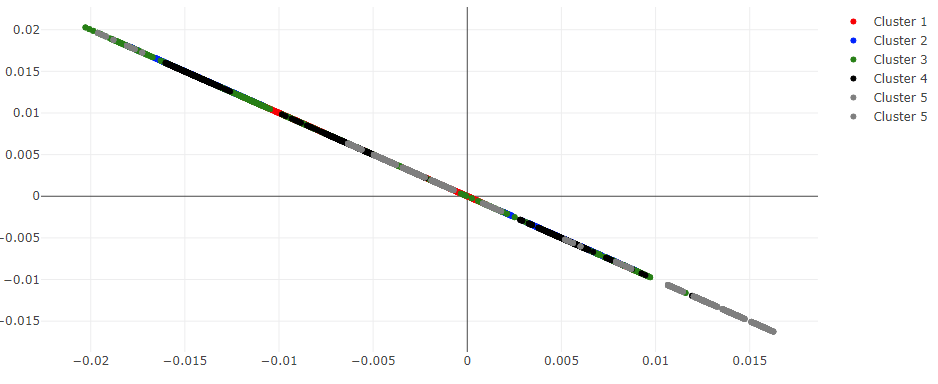
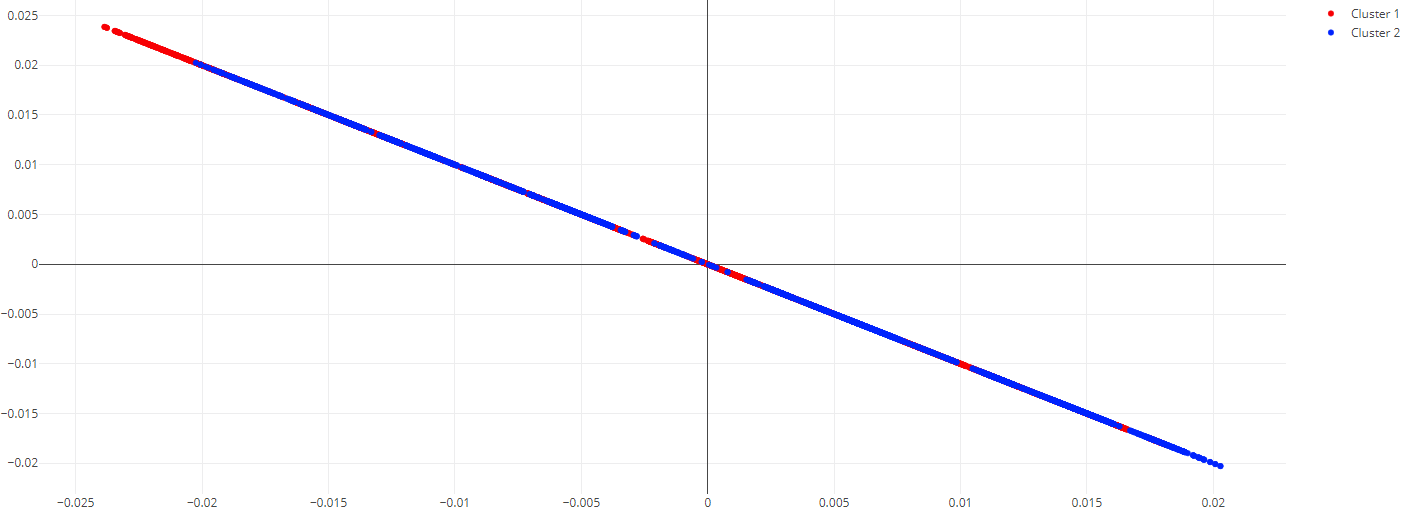


**ICA**

In the below graph adding more clusters did not change the shape of the distribution however the data is evenly clustered around centroids. I presume that the dimensions are changed by looking at how the graphs appear below. Looking at the axis we can see this.

Components: 2

Clusters:2 Clusters = 6

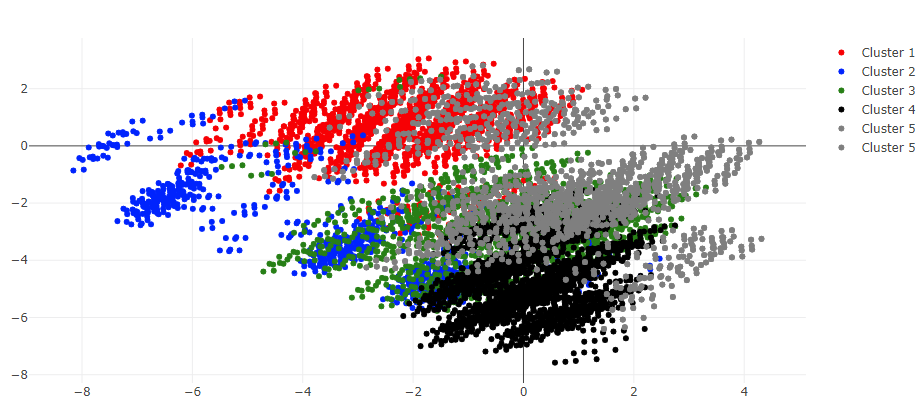
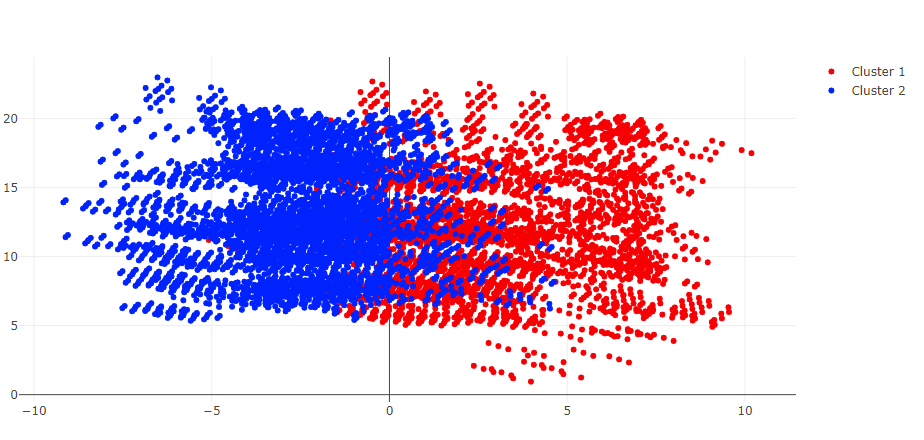


**Randomized Projection**

In the below graph adding more clusters change the shape of the distribution and the data is evenly clustered around centroids.

Components: 5

Clusters:2 Clustering= 6

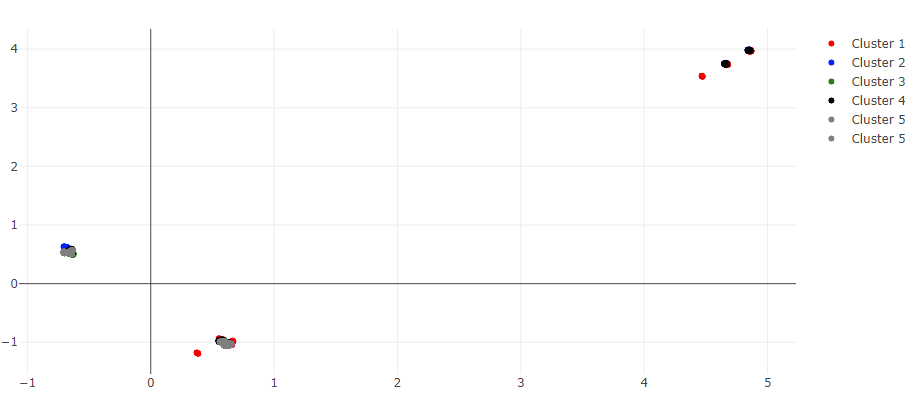
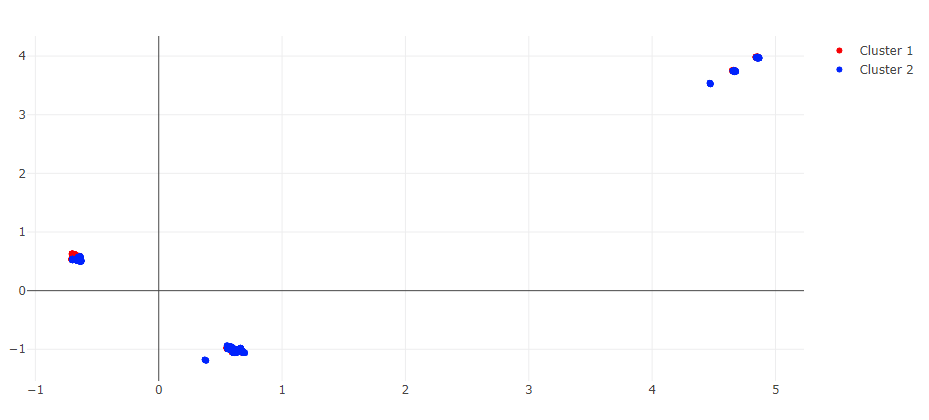


**Factor Analysis**

In the below graph adding more clusters did not change the shape of the distribution however the data is evenly clustered around centroids.

Components: 5

Clusters:2 Cluster= 6



**Sources**

Wikipedia contributors. (2019, March 14). Cluster analysis. In *Wikipedia, The Free Encyclopedia*. Retrieved 02:35, March 25, 2019, from <https://en.wikipedia.org/w/index.php?title=Cluster_analysis&oldid=887693442>

[Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Jin X., Han J. (2011) Expectation Maximization Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA