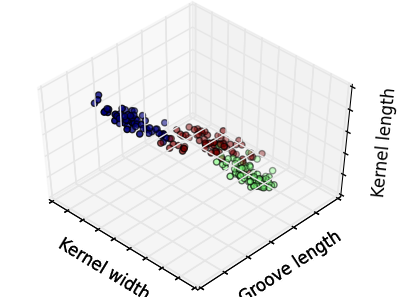
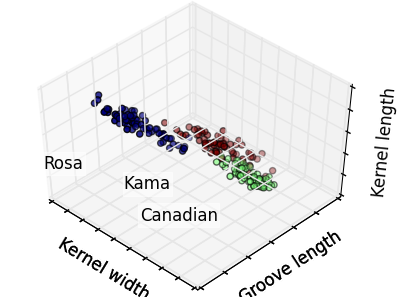
**AP4 Report**

The Seeds dataset from the Polish Academy of Sciences in Lublin contained measurements of geometrical properties of kernels belonging to three different varieties of wheat. A non-destructive soft X-ray technique was used to analyze the internal kernel structure and create a visual representation of the geometric parameters. The attributes that were measured include area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient, and length of kernel groove. These measurements can be used for classification and cluster analysis in order to predict which variety of wheat each kernel belongs to, including Kama, Rosa, and Canadian. A 3-cluster k-means system was constructed with the Scikit-Learn KMeans clustering algorithm using the seed data. The variety attribute was removed from the input data and was used instead to compare the success of the clustering. Accuracy was determined by comparing the k-means clusters with the actual seed varieties from the dataset. Each cluster was labeled with its respective variety, using the majority seed variety for any mislabeled inputs in a cluster. The accuracy was calculated using the built-in Scikit-Learn function, accuracy\_score, which computed a value of 89.52% accuracy from the clustering. The k-means cluster centroids were initialized by default to speed up convergence. The k-means algorithm assigns the data points to the closest centroid by finding the value that minimizes the distance between the data points and the cluster centroids. This is accomplished by calculating the sum of squared error for each data point in relation to the nearest cluster. The k-means algorithm selects the cluster with the smallest error. The centroids are then recalculated as the average of all data points in a cluster. These steps are repeated until the algorithm completes the maximum number of iterations of the k-means algorithm for a single run.



*3-cluster k-means system*



*Ground truth*