# Project 2

## Introduction

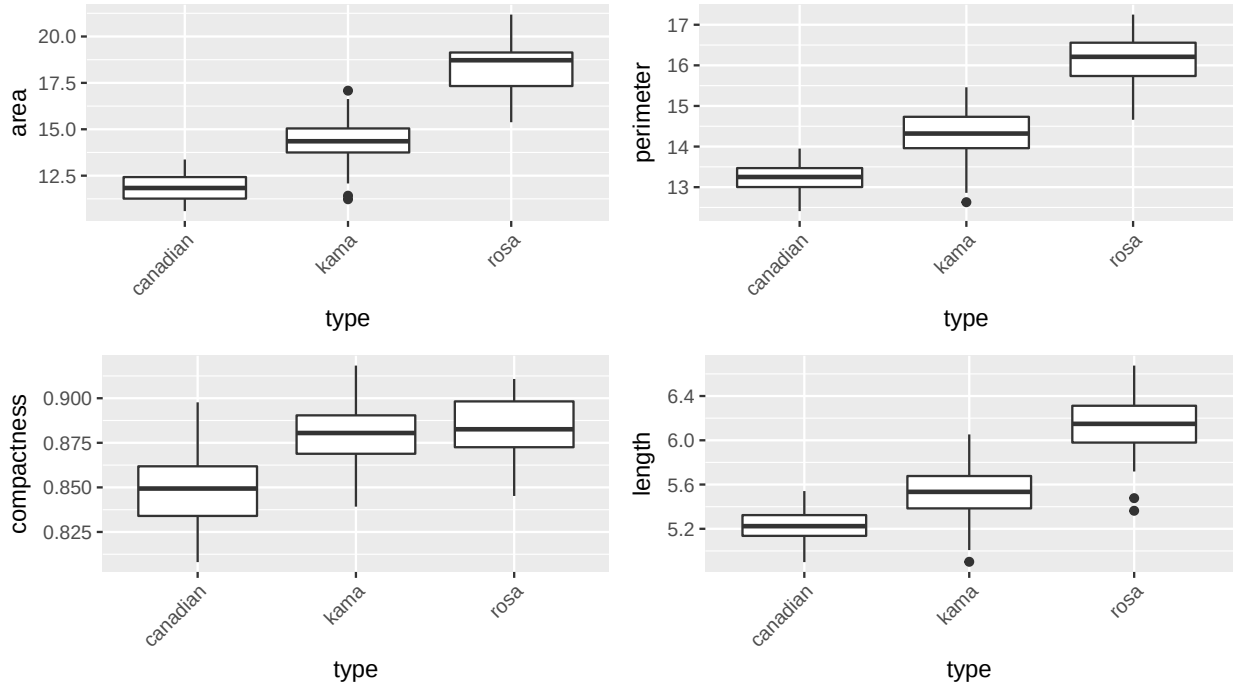
This is Project 2 for STAT 557 2018 Spring by Meridith Bartley and Fei Jiang. The aim of this project is to practice the k-means algorithm and the k-nearest neighbor algorithm. In this project we applied both algorithms to seed data in order to classify/cluster by seed type.

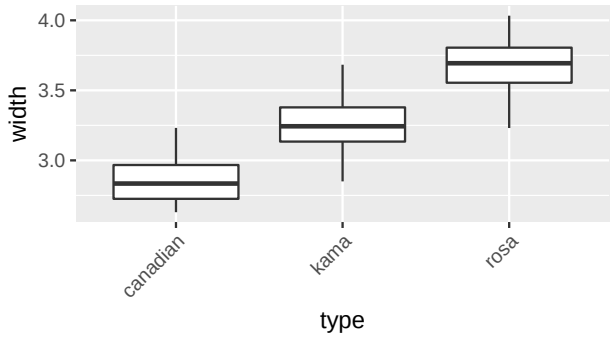
## Description of Data

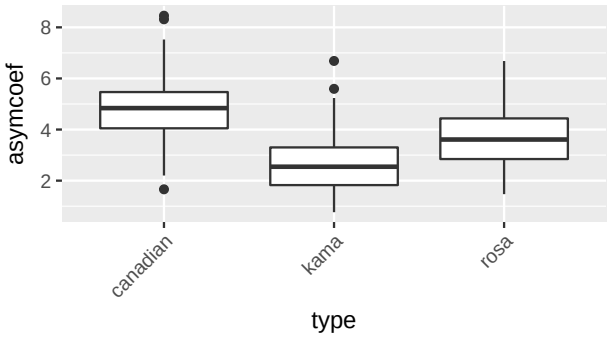
The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13x18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin.

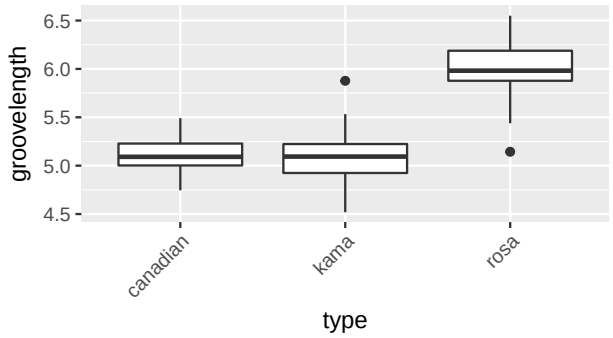
Boxplots for each attribute used as explanitory variables in the subsequent classification models are included below. This EDA allows for early indication of which variables may possibly be ommitted during dimention reduction. That is, what properties do not differ significantly between seed types.

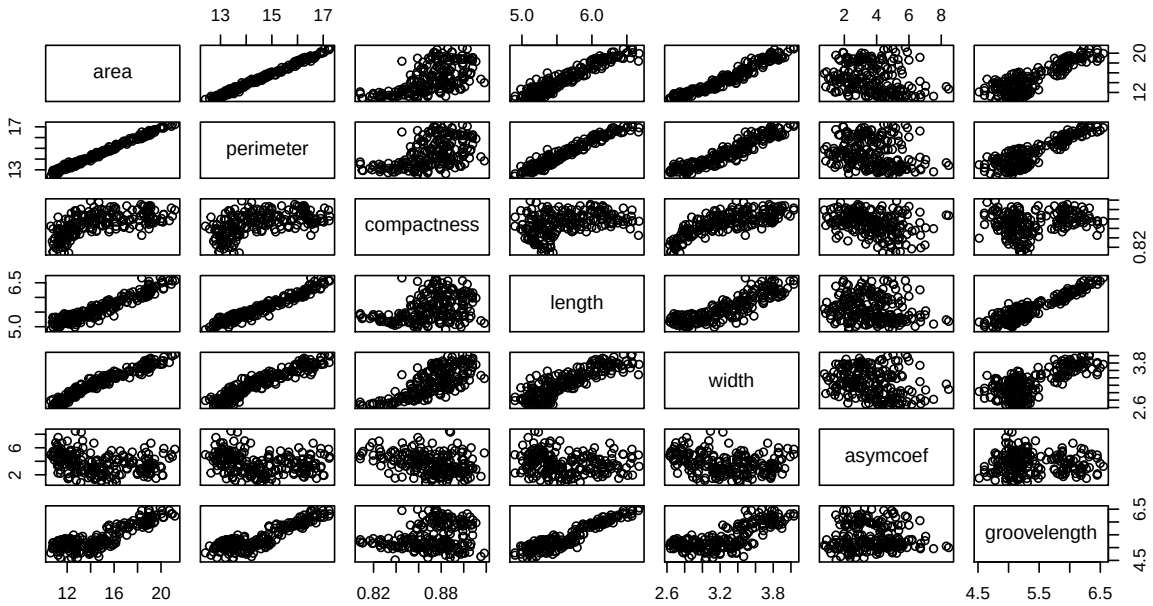
### Exploritory Data Analysis






## Principle Component Analysis

In order to test whether dimension reduction will improve predictions we also conducted Principle Component Analysis on the original dataset to get a new dataset with fewer dimensions. According to our PCA results, the first two component in total can explain about 99.302% of variance of the original database. The coefficients of the relevent componets are listed in the table below. Therefore, we took the first two components and the seed type values to build a new dataset with less dimensions.

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| Standard deviation | 3.28532 | 1.459265 | 0.2713485 | 0.1135231 | 0.0524235 | 0.0396289 | 0.0054457 |
| Proportion of Variance | 0.82939 | 0.163630 | 0.0056600 | 0.0009900 | 0.0002100 | 0.0001200 | 0.0000000 |
| Cumulative Proportion | 0.82939 | 0.993020 | 0.9986800 | 0.9996700 | 0.9998800 | 1.0000000 | 1.0000000 |

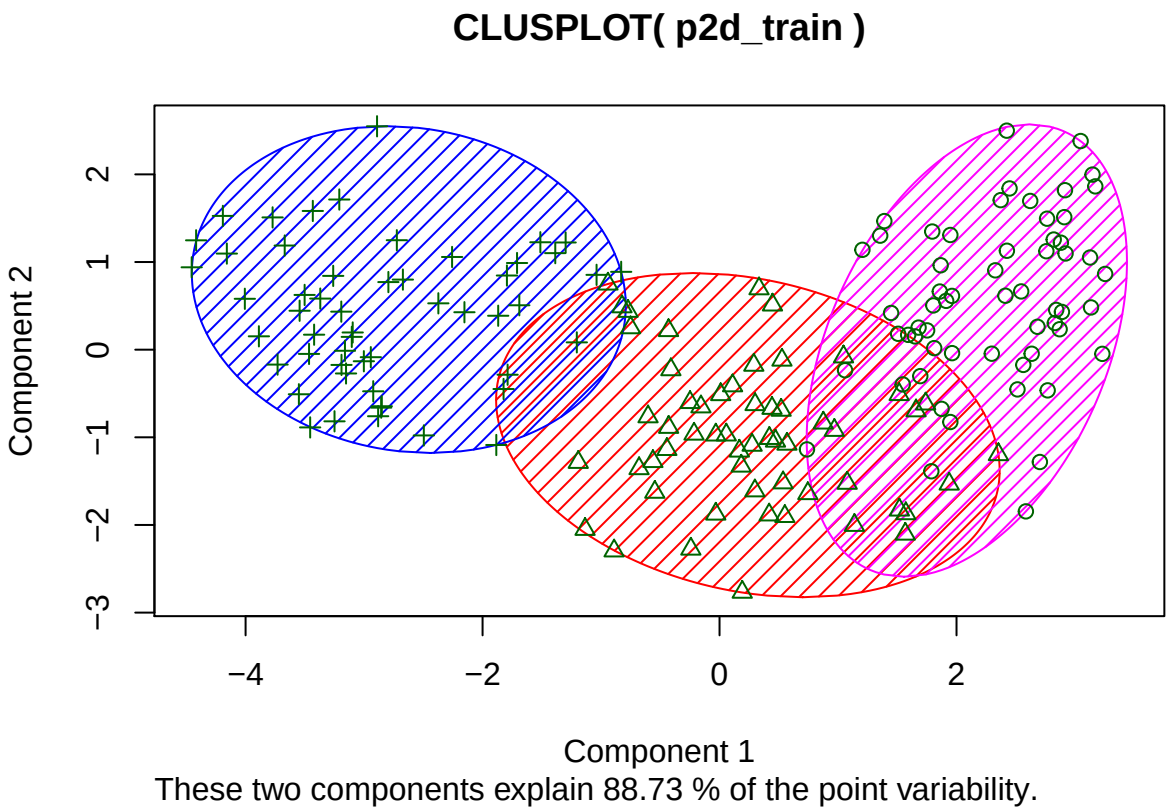
| | PC1 | PC2 |
|---|---|---|
| area | -0.8842285 | 0.1008058 |
| perimeter | -0.3954054 | 0.0564896 |
| compactness | -0.0043113 | -0.0028947 |
| length | -0.1285445 | 0.0306217 |
| width | -0.1110591 | 0.0023723 |
| asymcoef | 0.1276156 | 0.9894105 |
| groovelength | -0.1289665 | 0.0822334 |

# Analysis

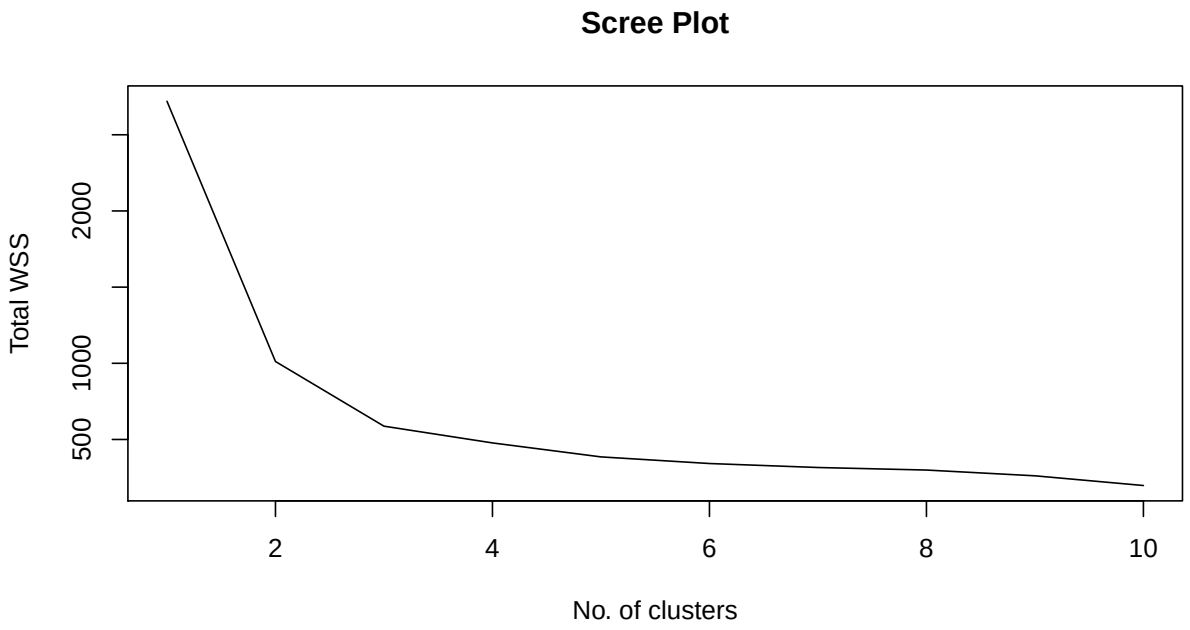
In the following analysis with two methods (k means and k-nearest neighbor algorithms - both supervised and unsupervised) and two datasets (original and dimension-reduced), we randomly selected 80% of the entire data as training data and the rest 20% as test data.

## K-Means Algorithm - Unsupervised clustering of Original Dataset

We initially conducted the k-means algorthm with on the original dataset and found that the overall prediction accuracy of our model in testing data is about 86%. In addition, we applied the k-means algorithem to a dimension reduced data set using the first two principal compoents and found that with the testing data there was about 86% accuracy. In the following plot we can see the cluster plot that uses PCA to draw the data using the first two principal components to explain the data.



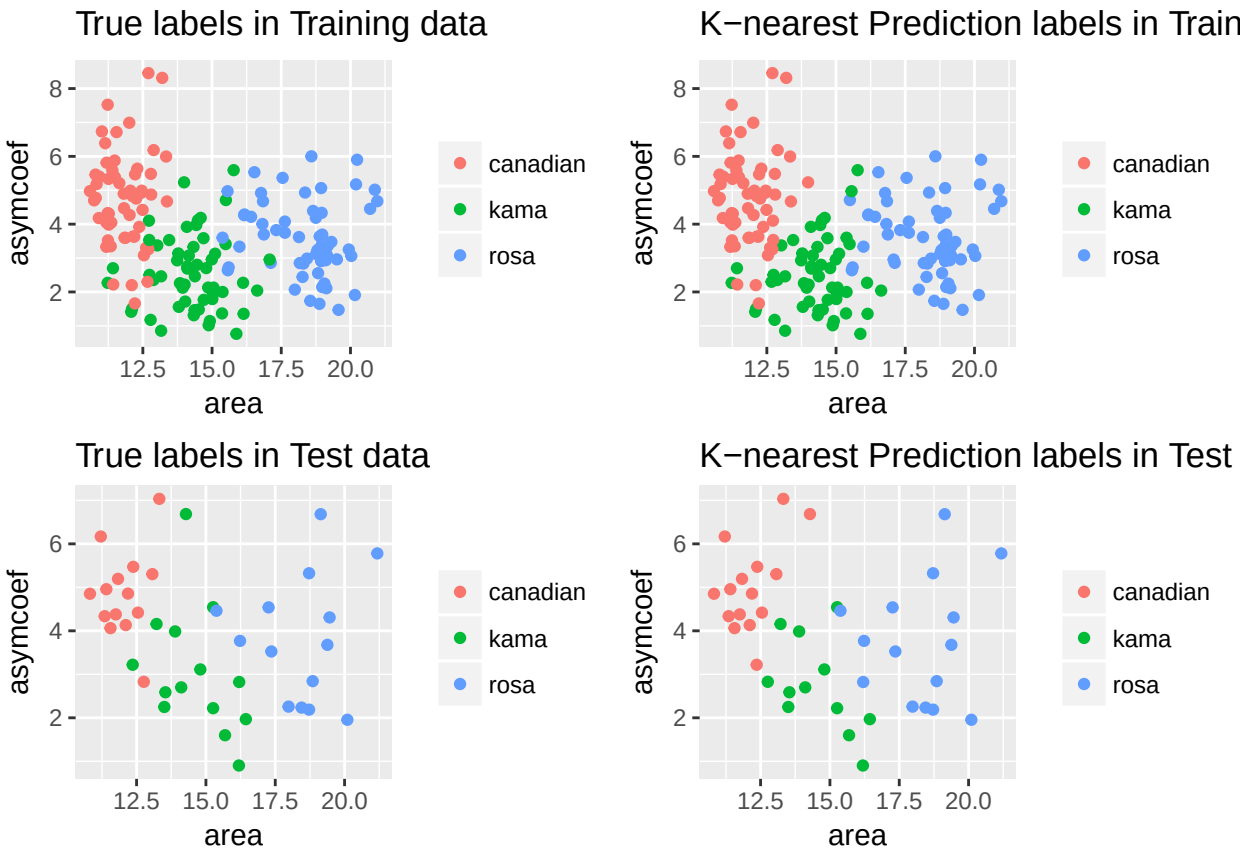

While we have reported the outcomes for k = 3, we also explored the possibilities of other size selections for k. We looked at total Within Groups Sums of Squares as a measure of how well our clusters fit the data. This is a measure of the distance the vectors in each cluster are from their respected centroid.The goal is to minimiza this value but no further than when the rate of improvement drops off. The scree plot below of these WSS values do indicate that is an appropraite number of clusters to choose. Indeed we do know from the data available that there are three seed types included.




## K-Nearest Neighbor - Supervised clustering in Original Dataset

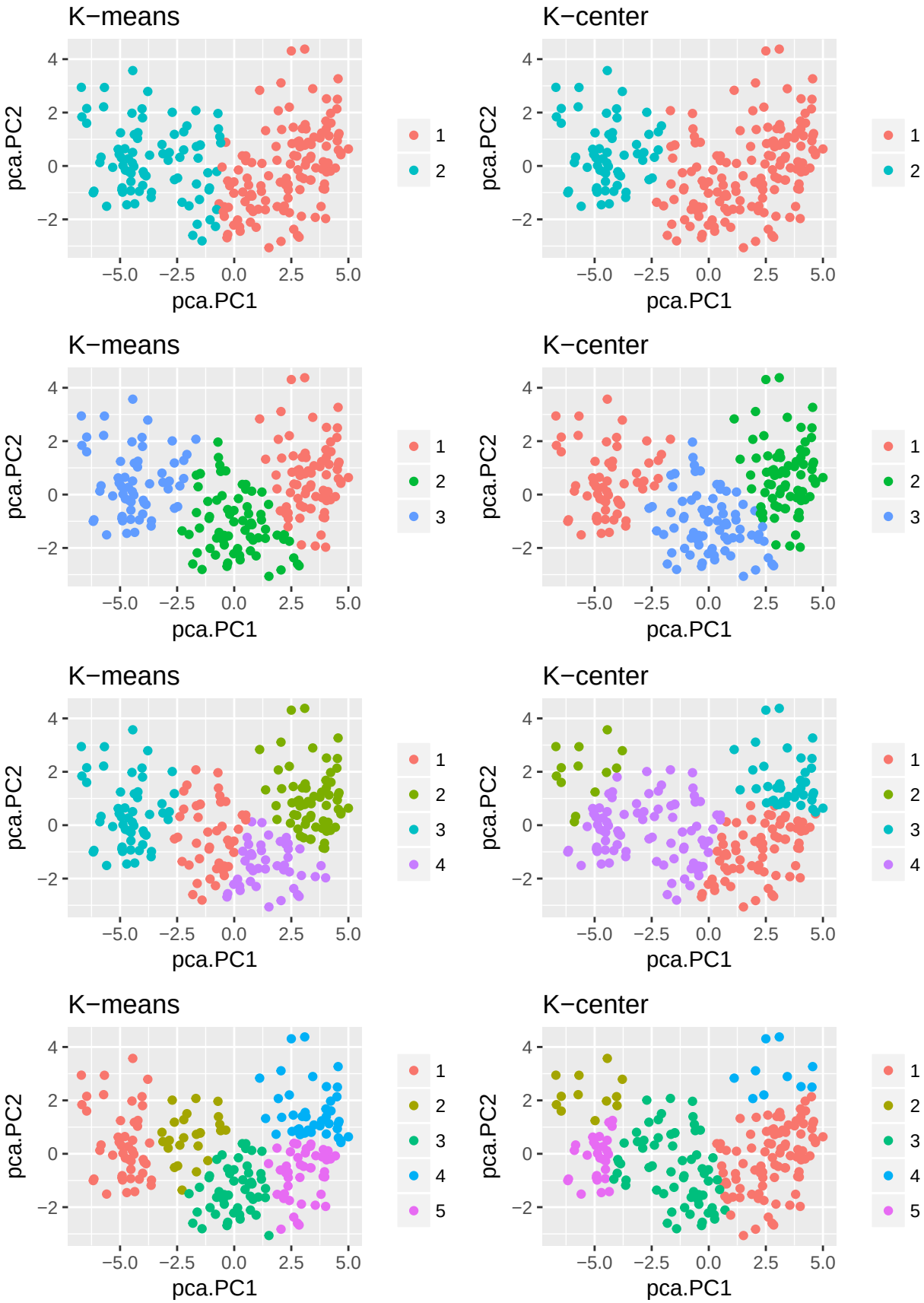
In this section, supervised K-nearest clustering is applied in the original dataset. The true and predictied lables in training and test data are shown in the figures below. In those figures, only first two predictors were shown. In general, in the training data, we obtained the accuracy rate of 92.9% and in the testing data, the accuarcy rate is about 90.5%. We think we reached a satisfying accuracy rate.

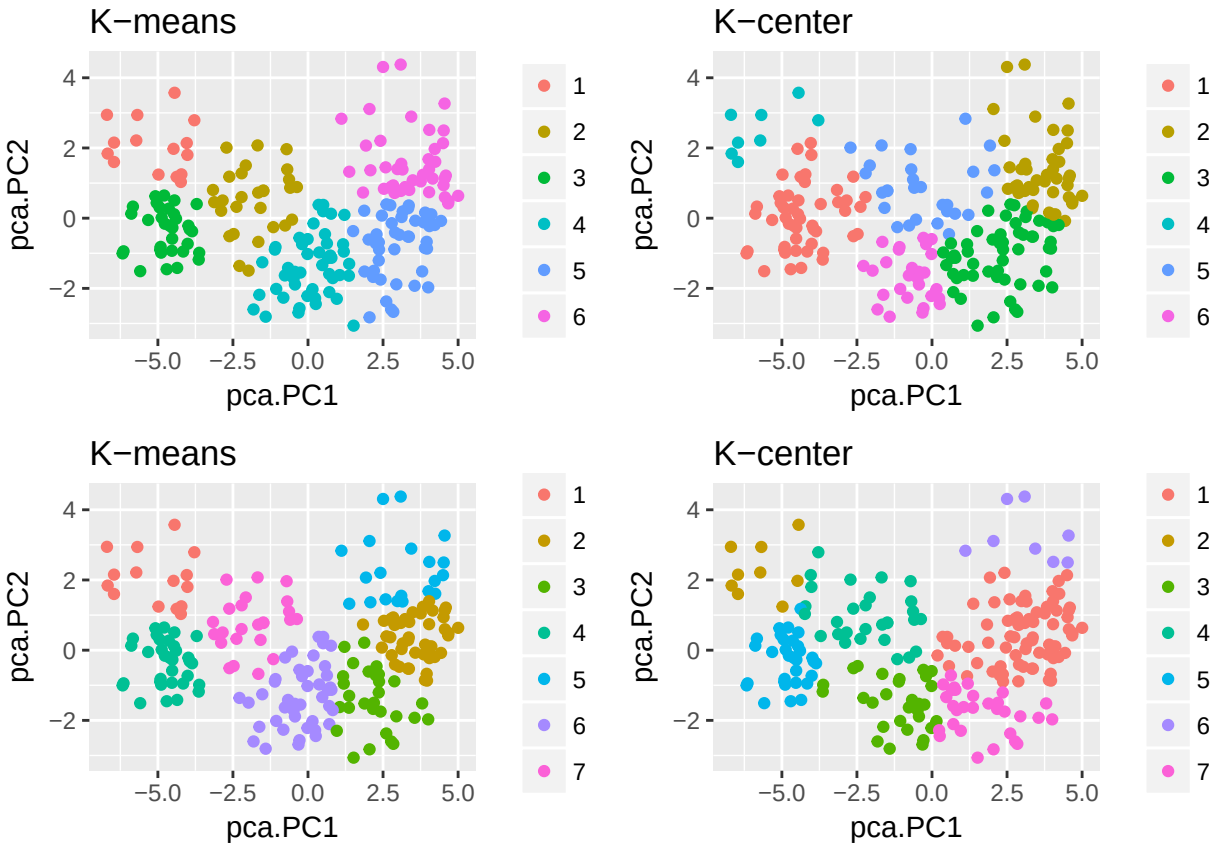
In the following plots we explore the true labels compared to those predicted by K-nearest neighbor clustering. We do this for both the original training and testing datasets. Not that we have plotted the data by area and the asymmetrical coefficent. It's clear that this method of clusing the data is performing well.

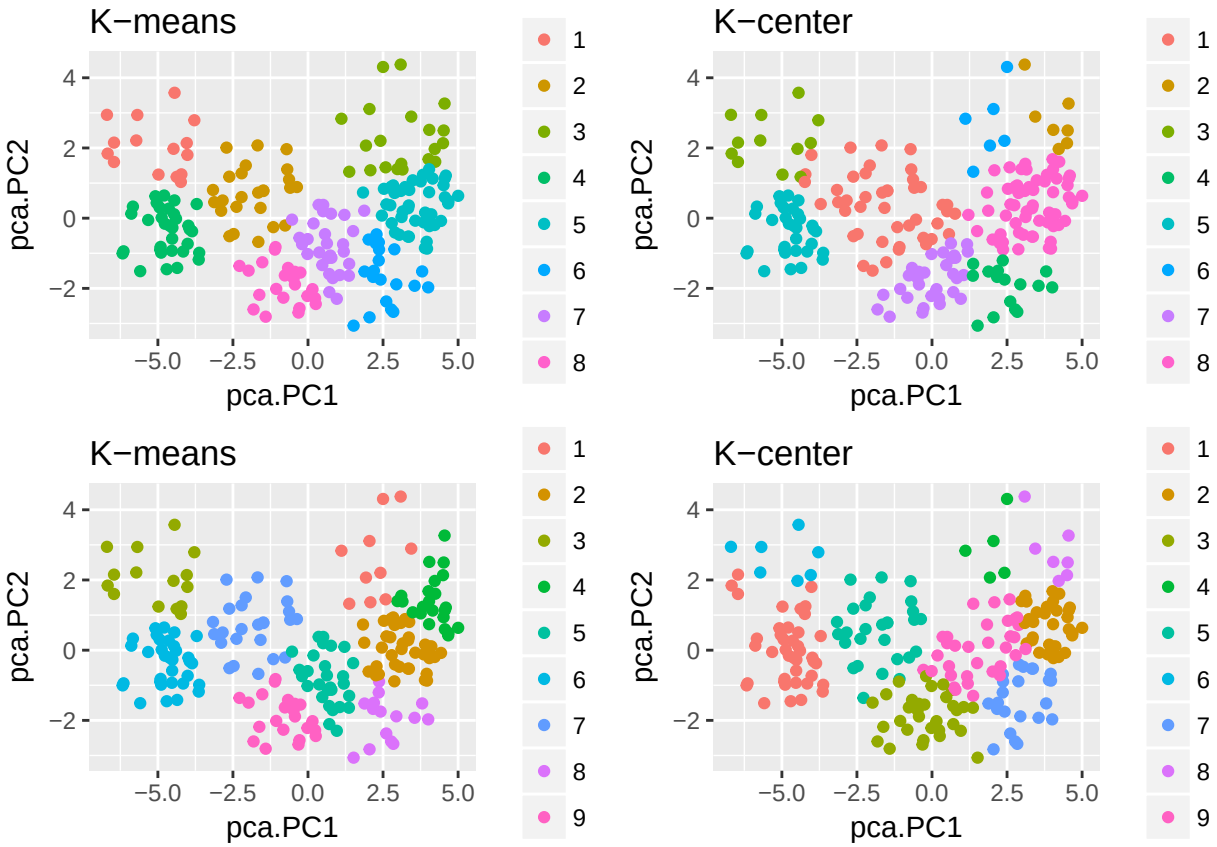


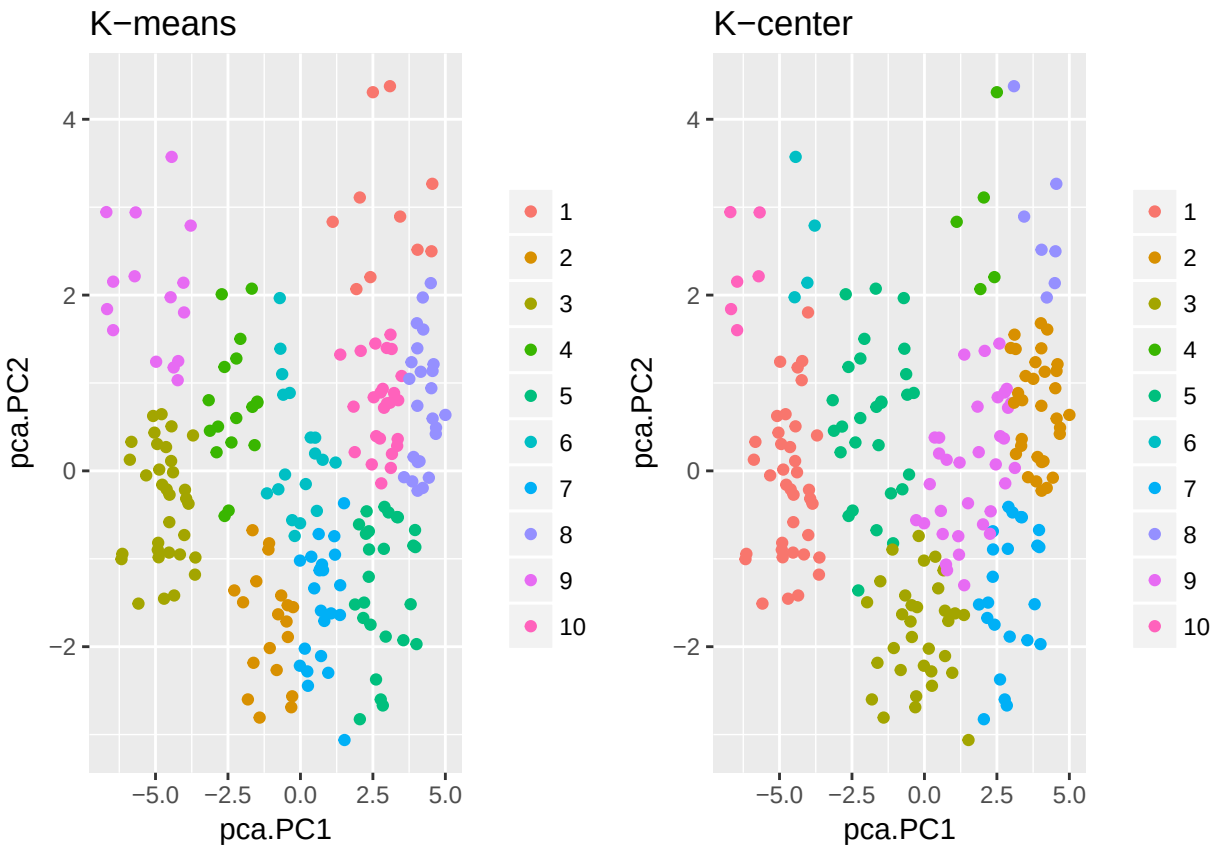

## K-Means and K-Centers Algorithm - Unsupervised clustering in Reduced Dataset (First Two Principle Components)

In this section, we applied unsupervised K-Means and K-Centers clustering algorithms to the reduced dataset (first two principle components).We tried 9 different numbers of clusters: 2 to 10 and plotted the scatter plot for each cluster number and for each algorithm. As the below figures show, there are differences between K-Center and K-Means clustering results. That is because k-means focuses on average distance while k-center focuses on worst scenario.


K−means
K−center
pca.PC1
pca.PC2
−5.0
−2.5
0.0
2.5
5.0
4
2
0
−2
1
2
3
4
5


K−means
K−center
K−means
K−center
pca.PC1
pca.PC2
4
2
0
−2
−5.0
−2.5
0.0
2.5
5.0
1
2
3
4
5
6
7


K−means
K−center
K−means
K−center
pca.PC1
pca.PC2
−5.0
−2.5
0.0
2.5
5.0
4
2
0
−2
1
2
3
4
5
6
7
8
9




## Conclusions

To sum up, we finisehd the following analysis in this project.

- The K-means clustering method on both the original and dimension reduced training and testing data, both showing similar accuracies of about 86% for the testing data.
- Confirmed a cluster size of by examining the Within Sums of Squares values.
- The supervised K-nearest clustering method achieved over 90% accuracy in both training and test data, which is satisfying.
- Applying differet k values, we explored the contrasting difference between k-means and k-center clustering methods.

## Contributions

The different tasks required to complete this project were equally divided between Meridith and Fei. K-means and cross-valiation analyses were completed by Meridith while Fei was responsible for K-nearest and K-center comparison. Both members of this group contributed to this report.