Project 1

# Introduction

This is Project 1 for STAT 557 2018 Spring by Meridith Bartley and Fei. The aim of this project is to practice linear classification methods and QDA and study basic techniques of dimension reduction.In this project we apply LDA, QDA, and multinomial logistic regression to soil sample data in order to classify into separate soil group (Orders).

# Description of Data

This dataset contains soil sample data over the US downloaded from Natural Resources Conservation Service (NRCS). After removing the incomplete data records, there are around 14,000 records left, each of which includes physical and chemical properties of soil samples (sand, silt, clay, organic carbon, bulk density, CEC soil, CEC clay, base saturation, and pH) and the corresponding soil classification group (soil order).

Boxplots for each physical and chemical property 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 soil Orders.

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(magrittr)  
library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(caret)

## Loading required package: lattice

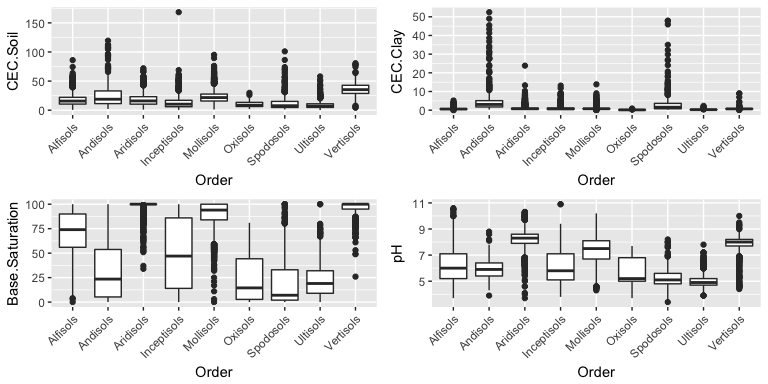
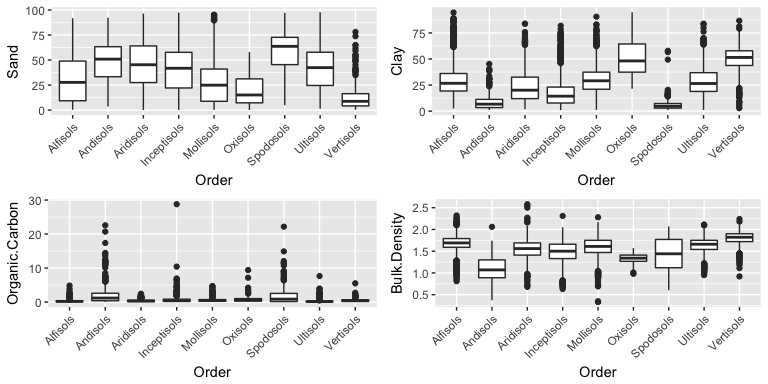
## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2018c.  
## 1.0/zoneinfo/America/New\_York'

library(nnet)   
library(scales)  
library(klaR)  
library(stats)  
library(grid)  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

## Exploritory Data Analysis



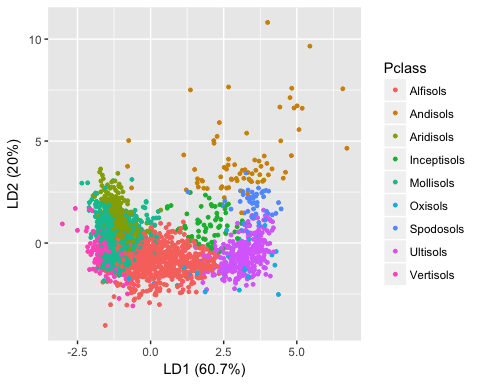
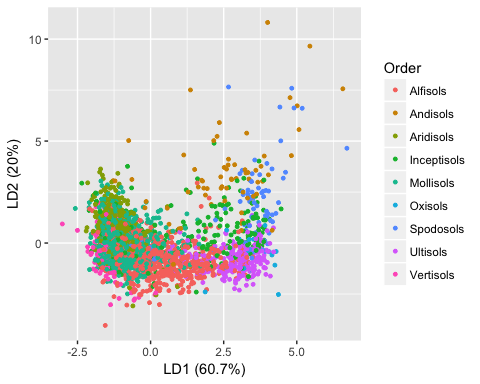
# Analysis

## Linear Discriminant Analysis (LDA) of original dataset (without dimension reduction)

#data partition  
train\_id <- caret::createDataPartition(y=project1data$Order, p=0.8,list = FALSE)  
#train <- project1data[lda\_train,]  
#test <- project1data[-lda\_train,]

## Alfisols Andisols Aridisols Inceptisols Mollisols Oxisols   
## 0.6230769 0.5675676 0.5552885 0.1698113 0.6265356 0.7647059   
## Spodosols Ultisols Vertisols   
## 0.2906977 0.7758621 0.5774648

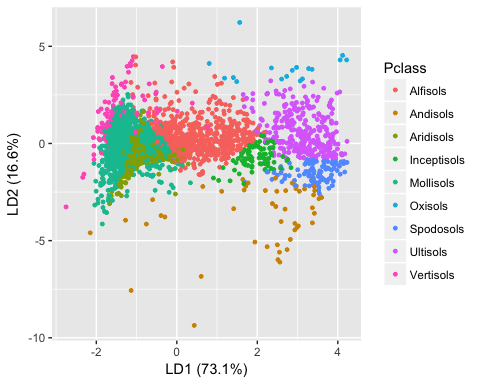
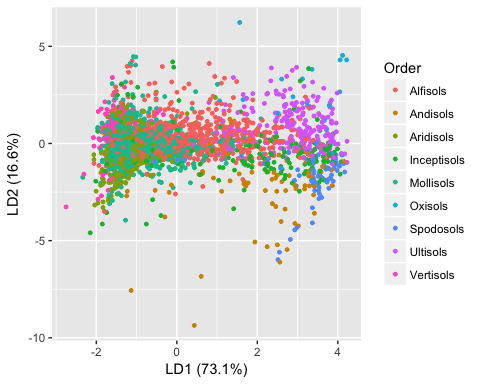
## [1] 0.5787955



## Linear Discriminant Analysis (LDA) of reduced dataset (with dimension reduction from PCA)

## Alfisols Andisols Aridisols Inceptisols Mollisols Oxisols   
## 0.4871795 0.3648649 0.4302885 0.1650943 0.6867322 0.4117647   
## Spodosols Ultisols Vertisols   
## 0.5581395 0.7974138 0.3591549

## [1] 0.5304724



## Quadratic Discriminant Analysis (QDA) of original dataset (without dimension reduction)

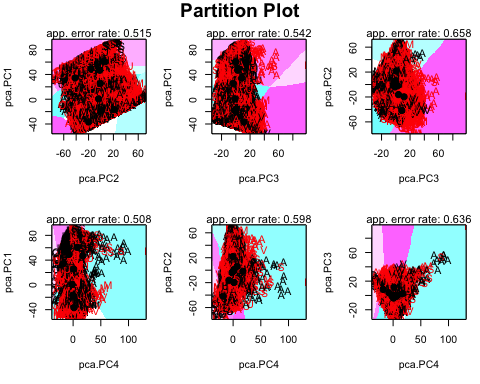
## Alfisols Andisols Aridisols Inceptisols Mollisols Oxisols   
## 0.7538462 0.4864865 0.5576923 0.1745283 0.4533170 0.8235294   
## Spodosols Ultisols Vertisols   
## 0.4534884 0.8060345 0.6760563

## [1] 0.5762712

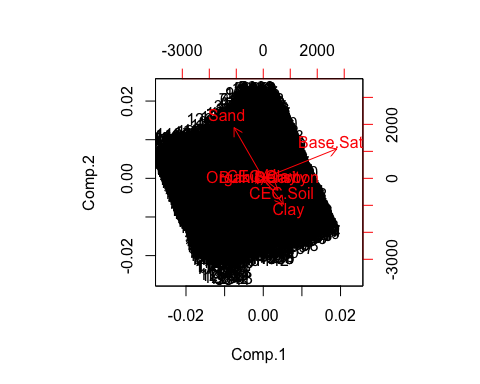
## Quadratic Discriminant Analysis (QDA) of reduced dataset (with dimension reduction from PCA)

## Alfisols Andisols Aridisols Inceptisols Mollisols Oxisols   
## 0.50769231 0.29729730 0.65625000 0.06603774 0.53439803 0.29411765   
## Spodosols Ultisols Vertisols   
## 0.63953488 0.78017241 0.57746479

## [1] 0.5275875



## Multinomial Logistic Regression



# Results

The results from these three approaches show that…

In order to compare the reults it is important to recall the diffences between these three classification approaches. The difference between LDA and logistic regression is that linear coefficients are estimated differently. MLE for logistic models and estimated mean and variance based on Gaussian assumptions for the LDA. LDA makes more restrictive Gaussian assumptions and therefore often expected to work better than logistic models IF they are met. QDA serves as a compromise between non-parametric methods (not explored in this project) and the linear LDA and logistic regression approaches. Since QDA assumes a quadratic decision boundary, it can accurately model a wider range of problems than can the linear methods. QDA can perform better in the presence of a limited number of training observations because it does make some assumptions about the form of the decision boundary.

# Contributions