

Project 1

Introduction

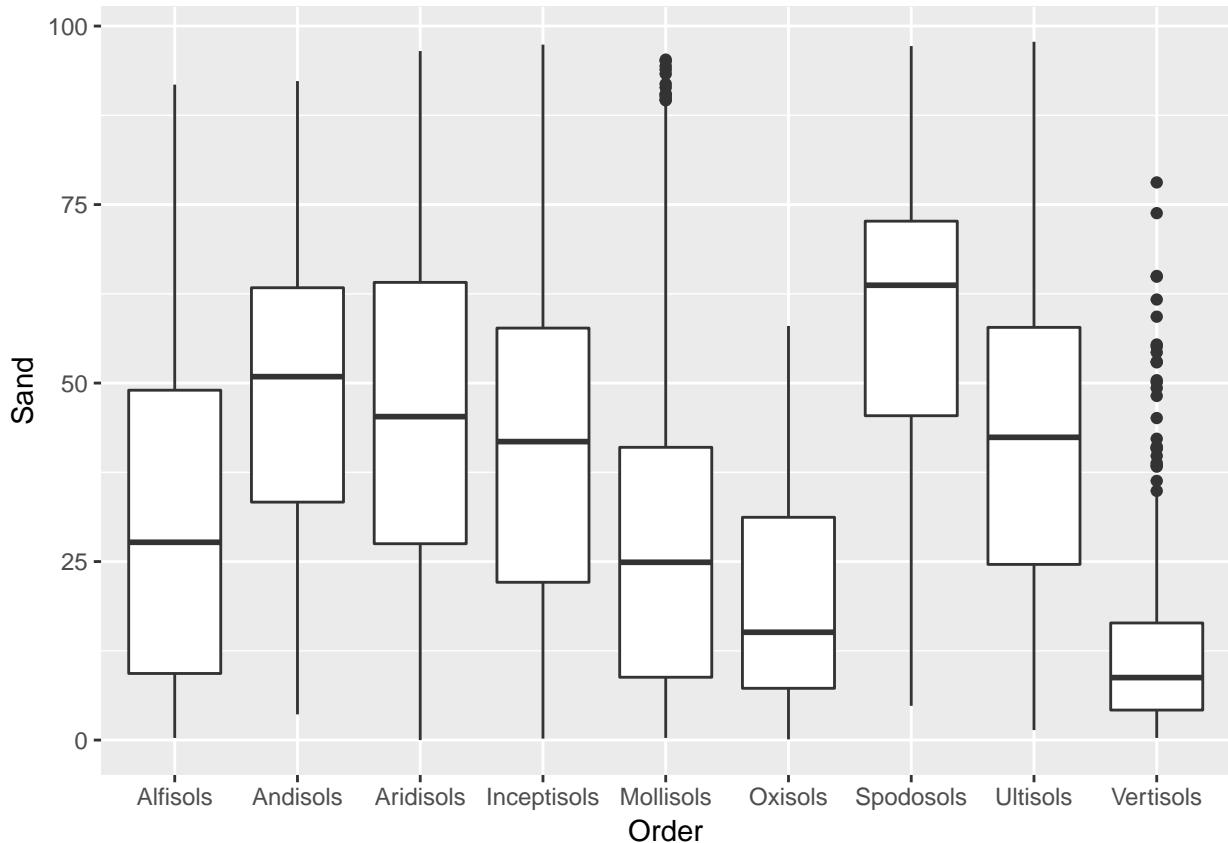
This is Project 1 for STAT 557 2018 Spring by Meridith Bartley and Fei. The aim of this project is to practice linear classification methods and QDA and study basic techniques of dimension reduction. In this project we (1) apply LDA, QDA, and multinomial logistic regression to soil sample data in order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

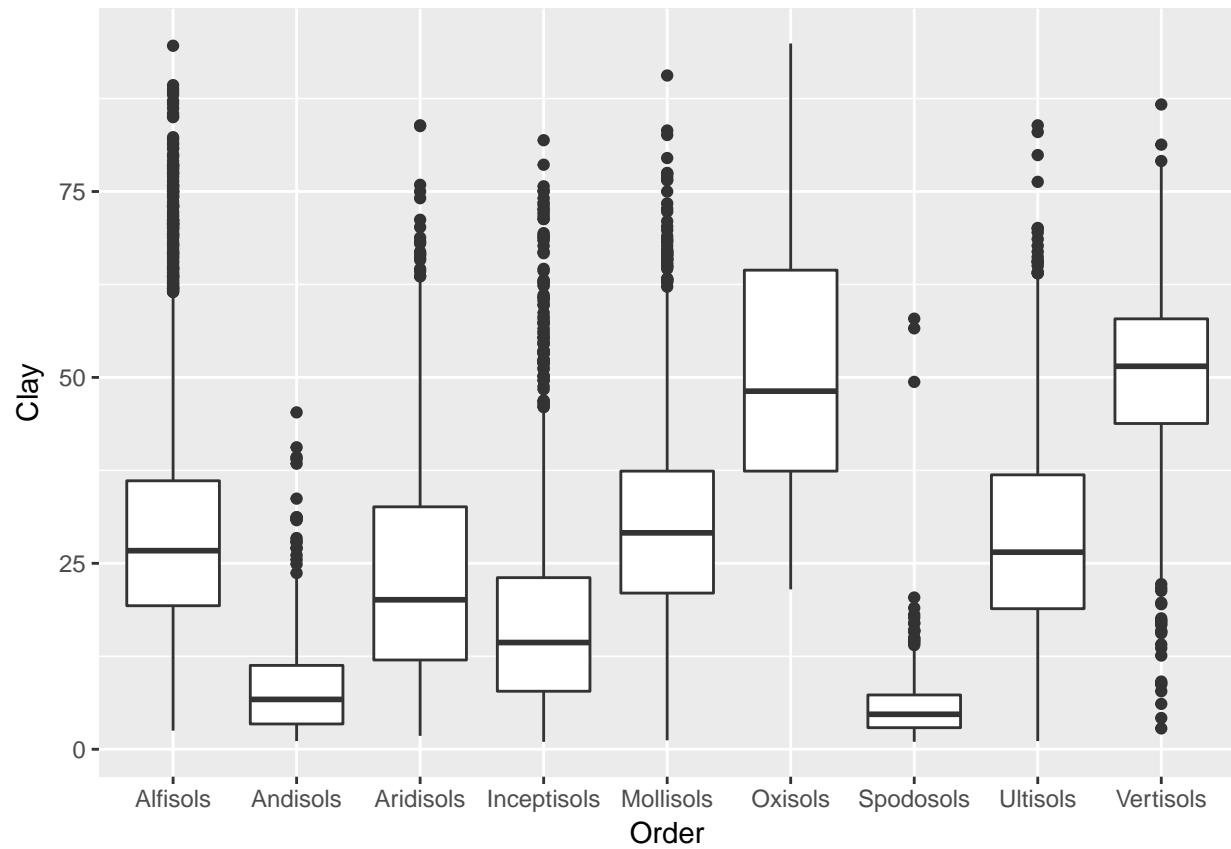
Description of Data

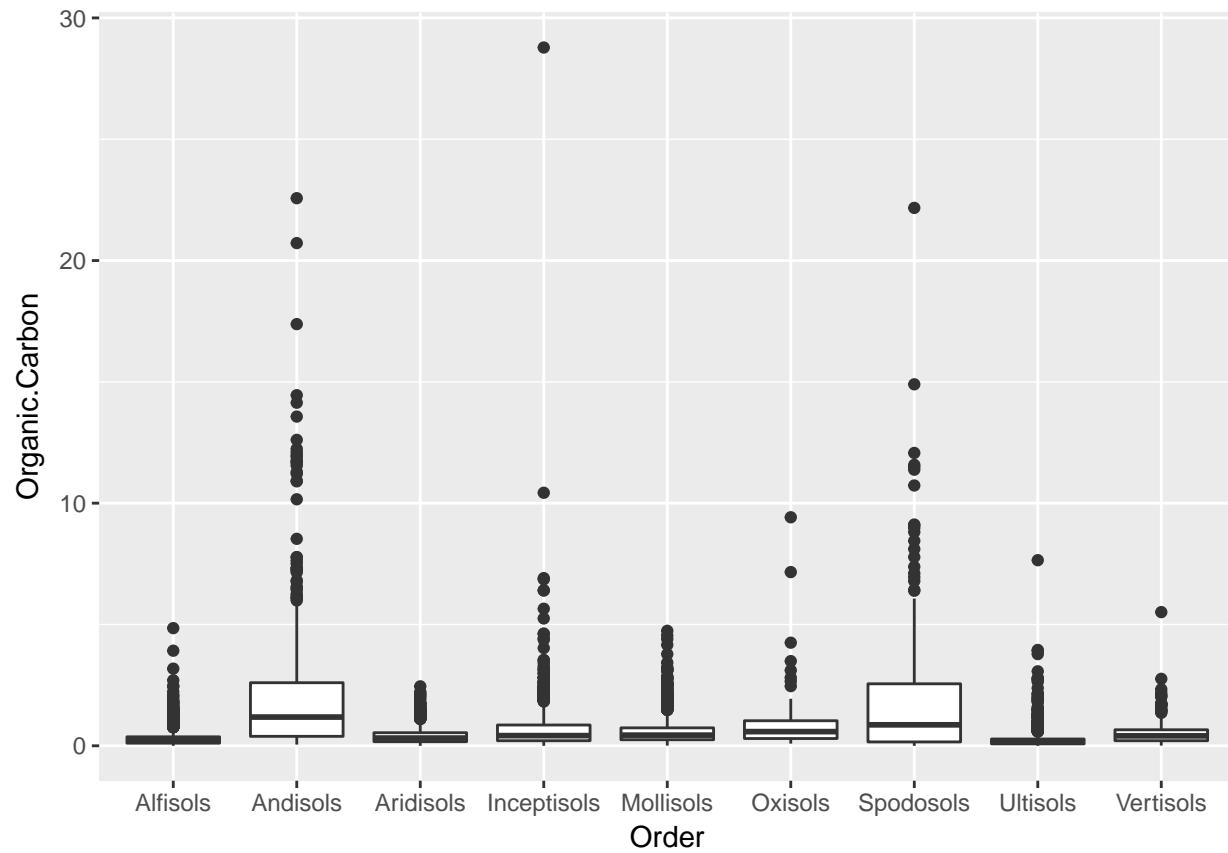
This is Fei's research data. It's about soil sample data over the US downloaded from NRCS, including the physical and chemical properties of soil samples (sand, silt, clay, organic carbon, bulk density, CEC soil, CEC clay, base satuaration, and pH) and the corresponding soil classification group (soil order).

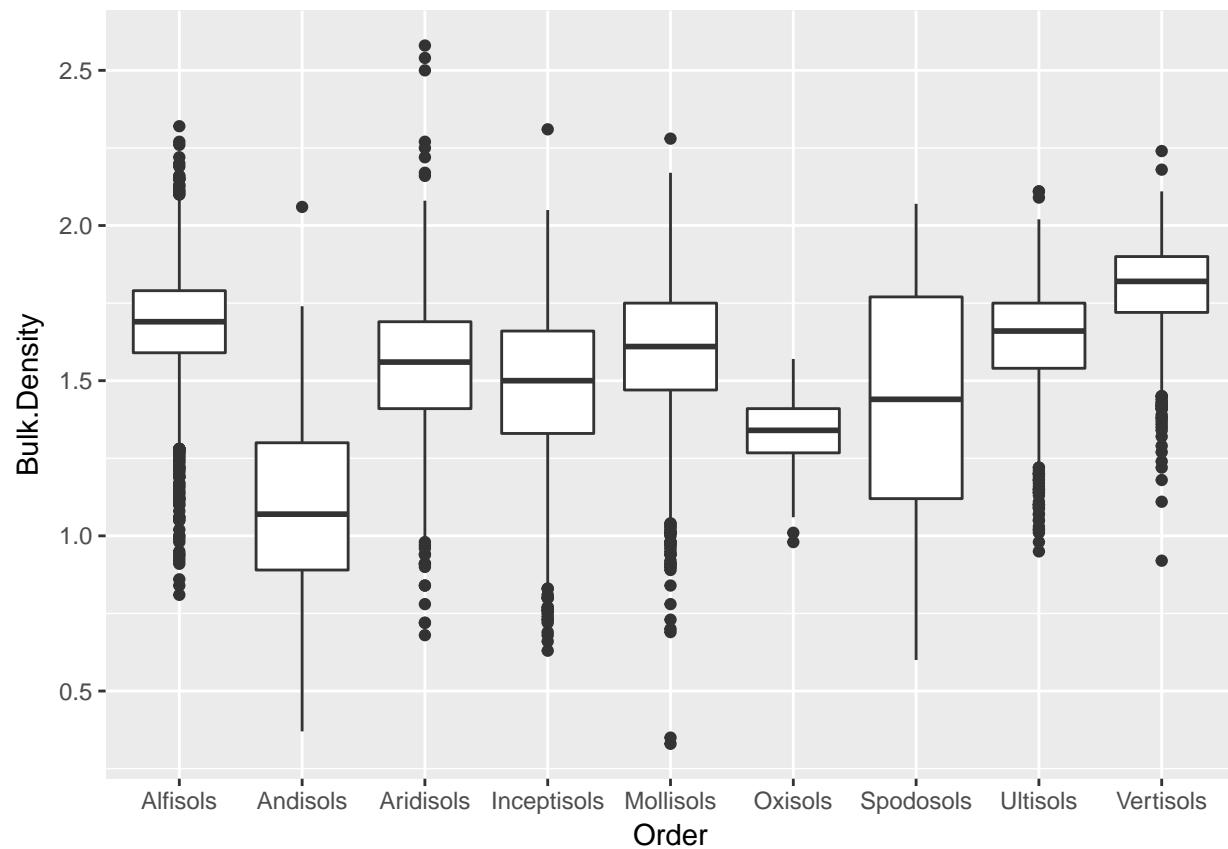
Boxplots for each physical and chemical property used as explanatory variables in the subsequent classification models are included below. This EDA allows for early indication of which variables may possibly be ommited during dimention reduction. That is, what properties do not differ significantly between soil Orders.

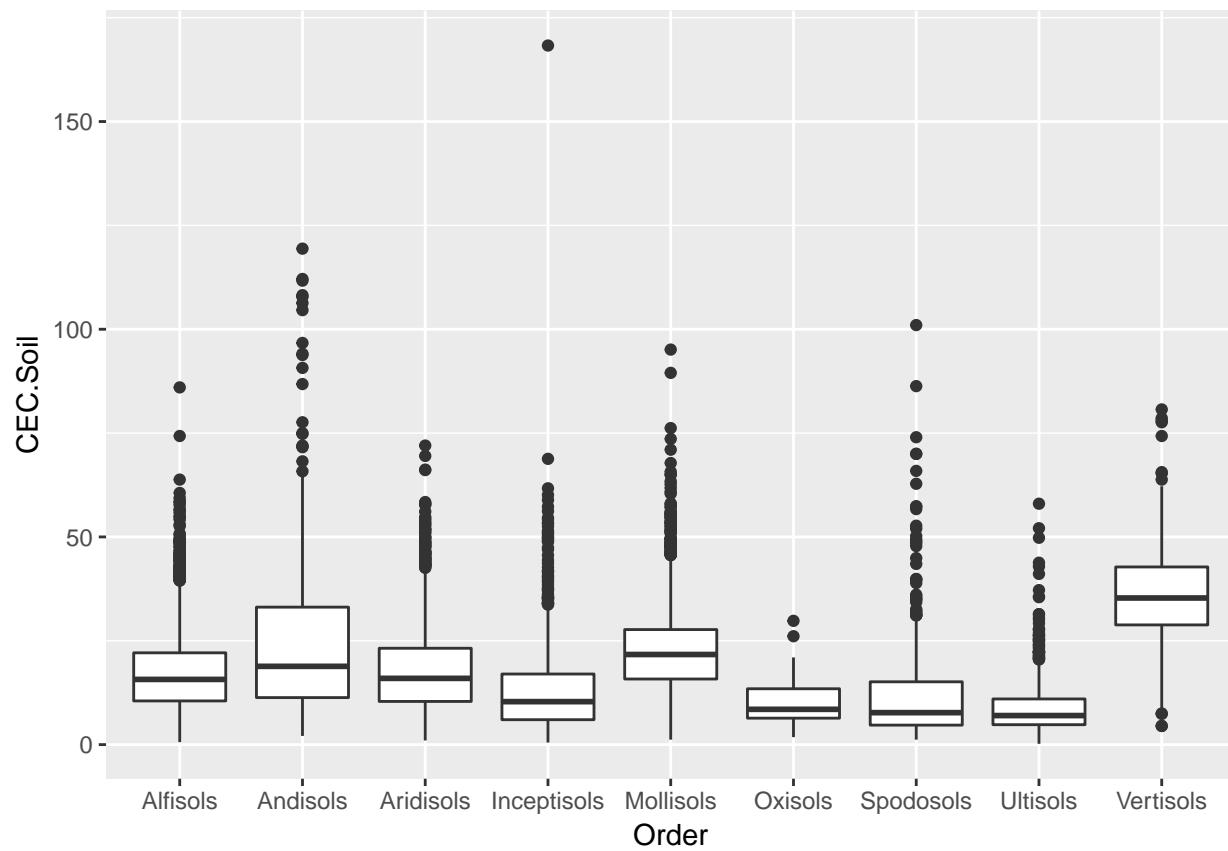
Exploratory Data Analysis

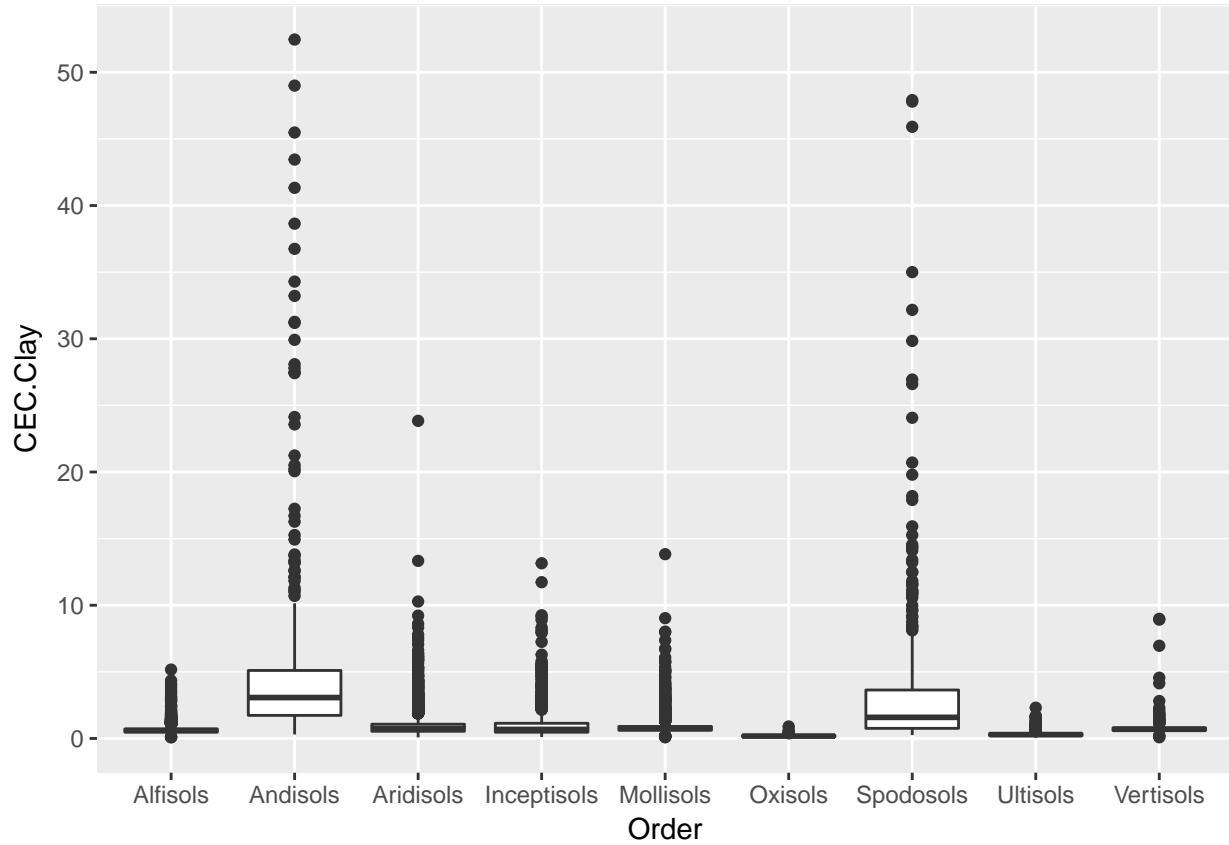


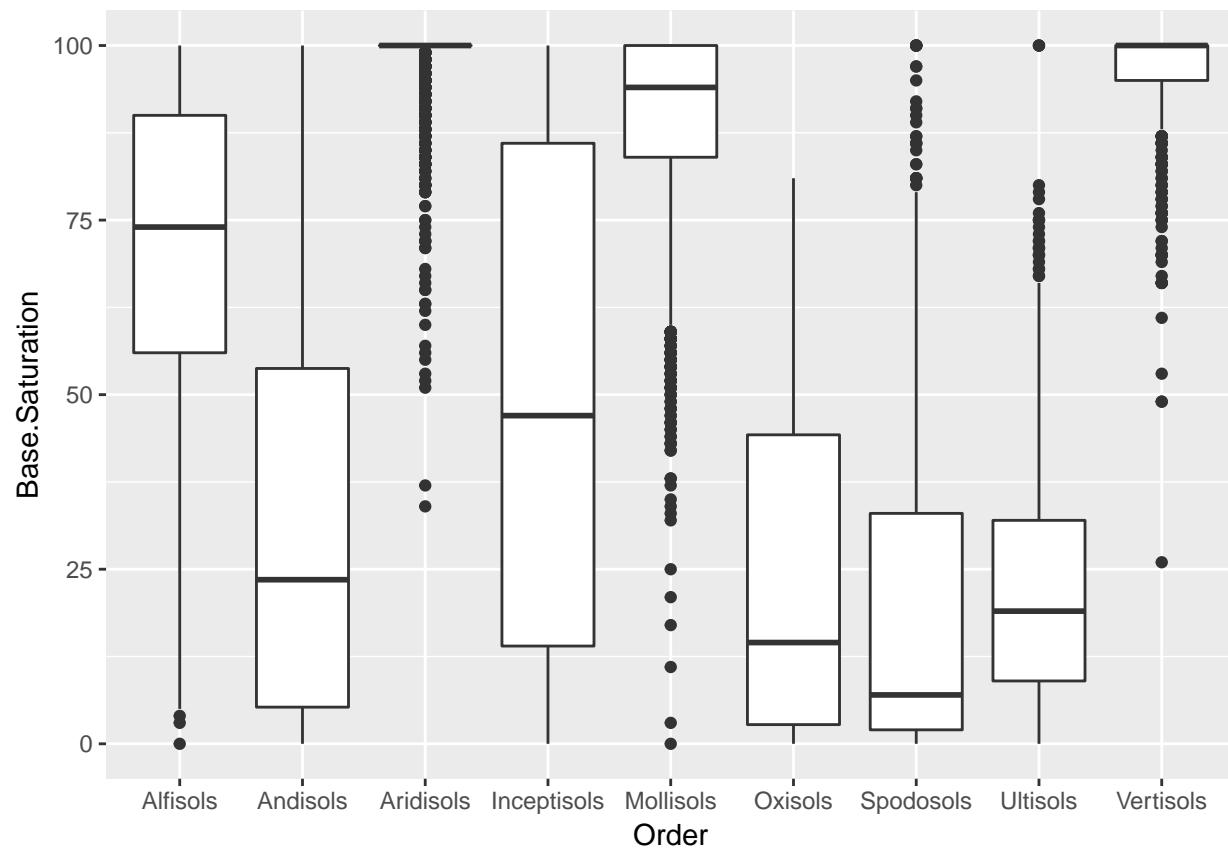


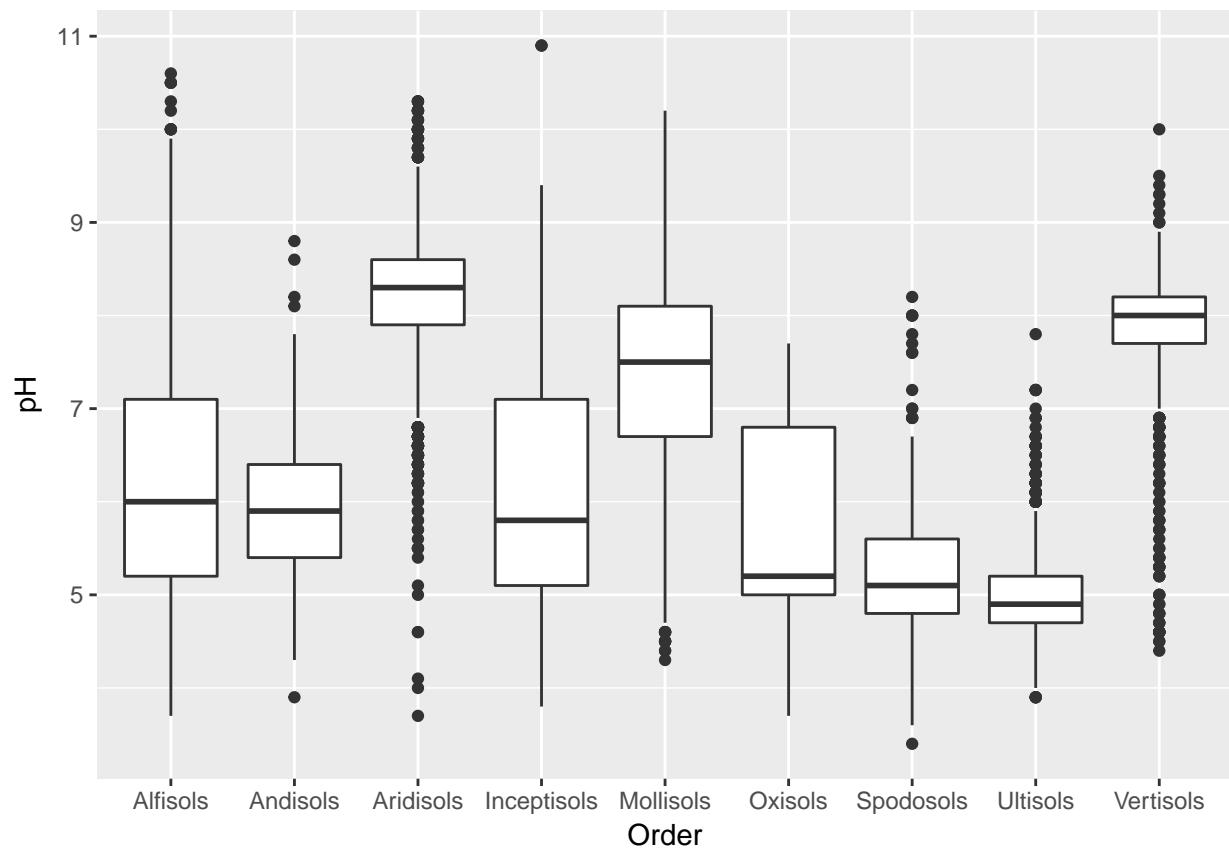


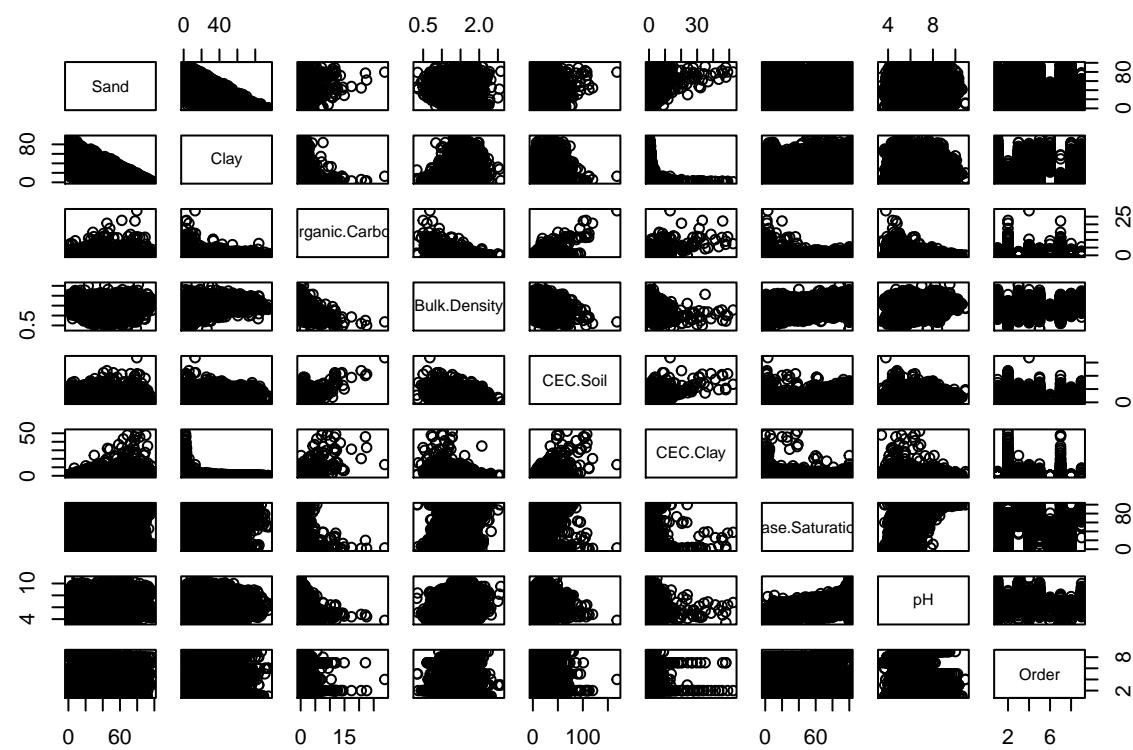






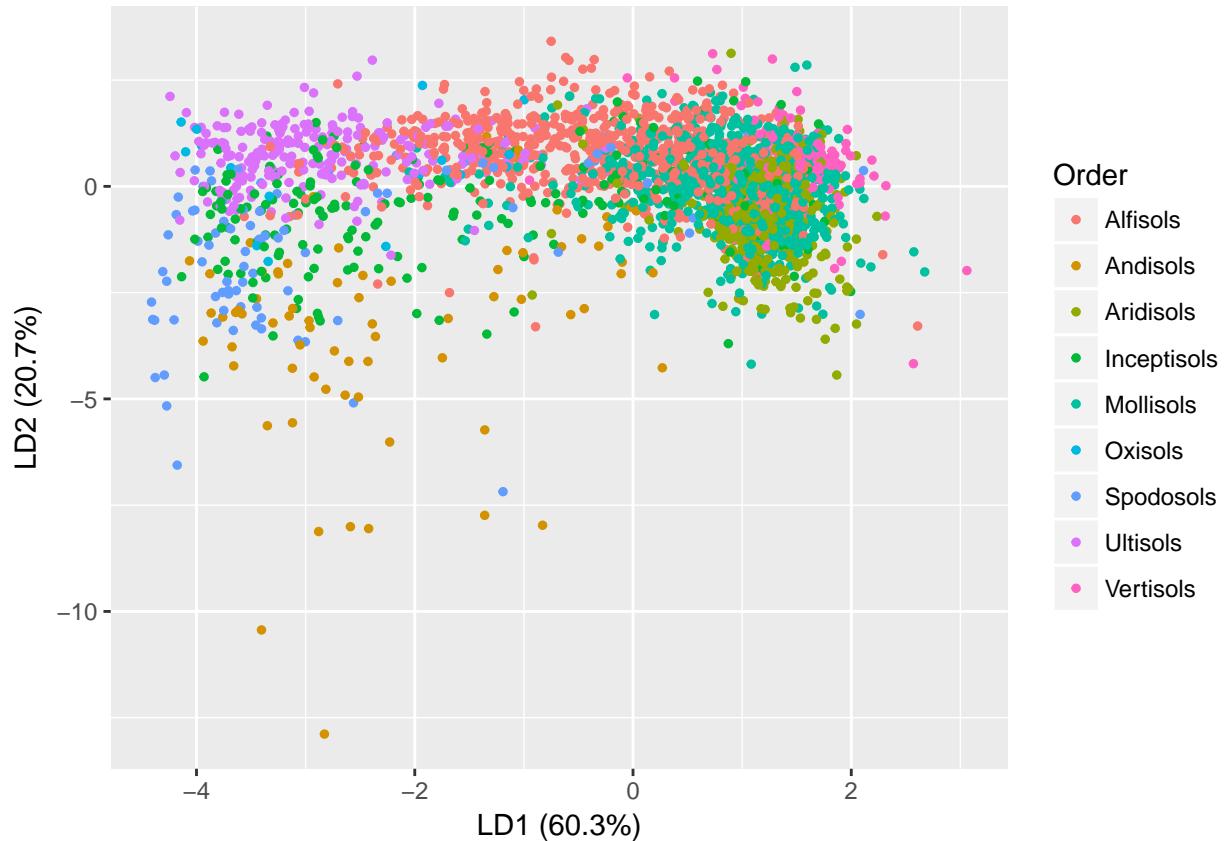


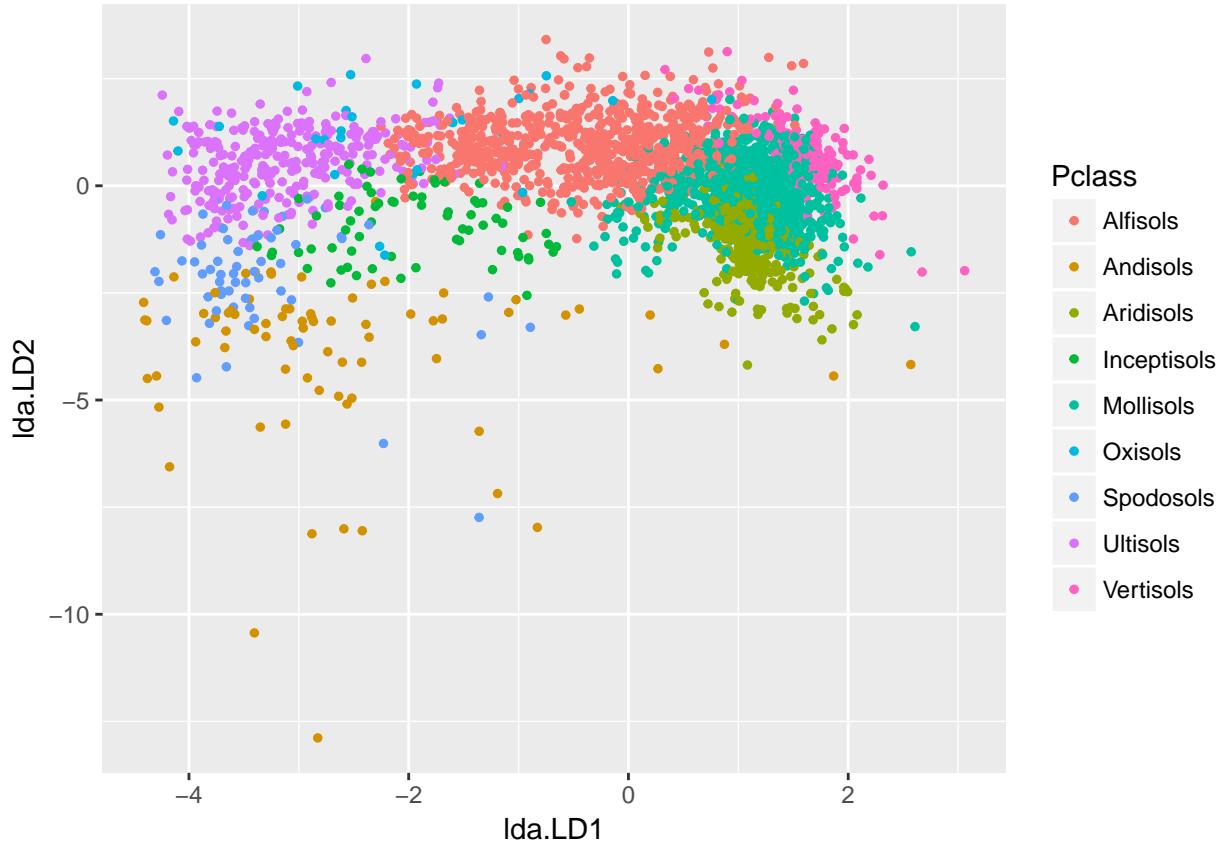




Analysis

Linear Discriminant Analysis (LDA)





Quadratic Discriminant Analysis

Multinomial Logistic Regression

Results

The results from these three approaches show that...

In order to compare the results it is important to recall the differences 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