Exploring Seeds and Automobile Datasets Using Statistical Methods

Ryan Smith
Statistics Department
University of California, Davis

***Abstract*- An exploration into the features of three varieties of wheat (Kama, Rosa, and Canadian), and how their variety can be determined by on other features using various classification methods. Also, an analysis into automobiles and what maximizes their price and horsepower, and a comparison of prediction methods.**

I. Introduction

In this report I will analyze two different data sets: one which contains information about three types of seeds, and another which contains information about various features of automobiles sold during the year 1985. Both of the data sets that I will perform analysis on are publicly available from the University of California, Irvine Machine Learning Repository.

A. Seeds dataset

The seeds data set that I will explore contains 210 seeds, 70 of each of the three types, with 7 measurements for each seed: area, perimeter, compactness, kernel length, kernel width, asymmetry coefficient, and kernel groove length. In this paper the focus will be on the type of seed and analyzing the best methods of classifying a new seed to its corresponding type. To do this I will first perform basic data exploratory data analysis to see the relationships between seed features. I will then perform data clustering to visualize the data and to see how distinct each seed type is. Finally, I will use several methods of classification to find the best method of classifying a seed for which the type is unknown.

B. Automobile dataset

The seeds data set that I will explore contains 193 automobiles, with 25 features for each car including price, highway miles per gallon, weight, and number of cylinders. In this paper I will focus on predicting two variables: price and highway miles per gallon. To accomplish this, I will first perform basic exploratory data analysis to get an understanding of the correlations between the variables in the dataset. I will then split the data into testing a training data repeatedly and use various prediction methods to find which minimizes the prediction error on unseen data. I will reduce the dimensions of the model by removing insignificant variables and attempt to minimize the error. To conclude, I will use Principal Component Analysis on the data set to reduce the number of dimensions of the data set while retaining a significant proportion of the original variance contained in the data set.

II. Analyzing the Seeds Dataset

A. *Exploratory Data Analysis*

To begin the exploration of the Seeds dataset, I will first plot a correlogram that will show the correlation between the variables in the dataset.

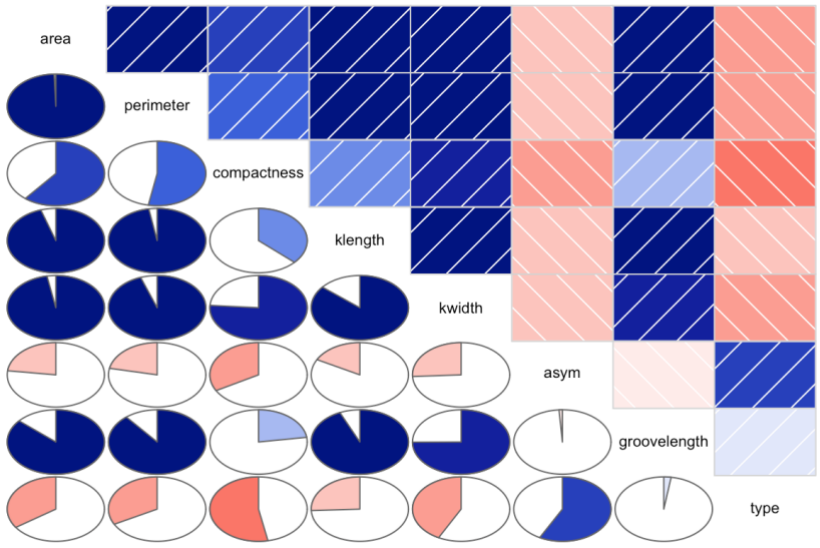

Figure 1. Correlogram of Seeds dataset. Blue represents a positive correlation, while red represents a negative correlation. Darker is stronger correlation, whereas lighter is a weaker correlation.

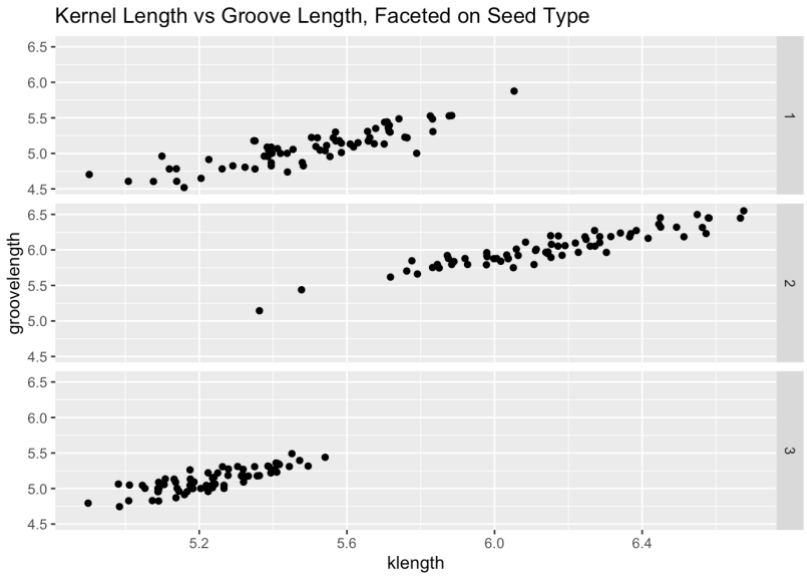

Figure 2. A scatter plot of groove length versus the length of the kernel, faceted on the type of seed.

Fig 1 shows the correlations between all eight variables in the data set. Several variables are highly correlated such as area and perimeter, which makes sense as area and perimeter are calculated using formulas containing the same variables. The variables which are most correlated with the type of seed are compactness, asymmetry

coefficient, and the width of the kernel. Fig 2 shows that the relationships between groove length and kernel length are distinct when compared across the three types of seeds. This distinctness can help classification of the seed type be more accurate. I will now plot the three variables that are the most correlated with seed type and distinguish the type of seeds by colors.

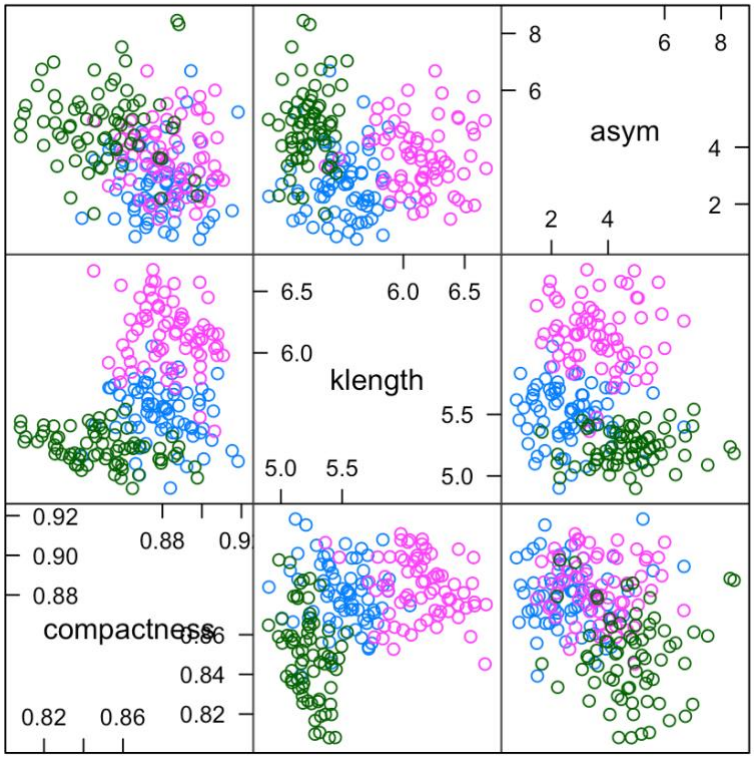

Figure 3. A scatter plot matrix of compactness, kernel length, and asymmetry. Pink is seed type 1, green is seed type 2, and blue is seed type 3.

Fig 3 highlights the distinctness of the three seed types. The clusters of each type are well separated according to the seed type(color), which implies that when I perform classification on the data set it should be accurate.

B. Clustering

I will now perform clustering on the Seeds dataset to separate the data into clusters. To do this I will use K-means clustering, which is a clustering method which partitions the data into k groups based on the distance of each data point to the means of each cluster iteratively. Each new data point is assigned to the cluster who's mean is closest to it. To perform K-means clustering, I need to specify the number of clusters to divide the data into. Fig 4 shows the sum of the squared distances from each data point in each cluster to its cluster mean. The more clusters you have, the smaller this distance will be in general for small quantities of K, but for this case I will choose 3 clusters as that is the number of seed types, we are interested in. T clustering will show how well the data fits into 3 groups.

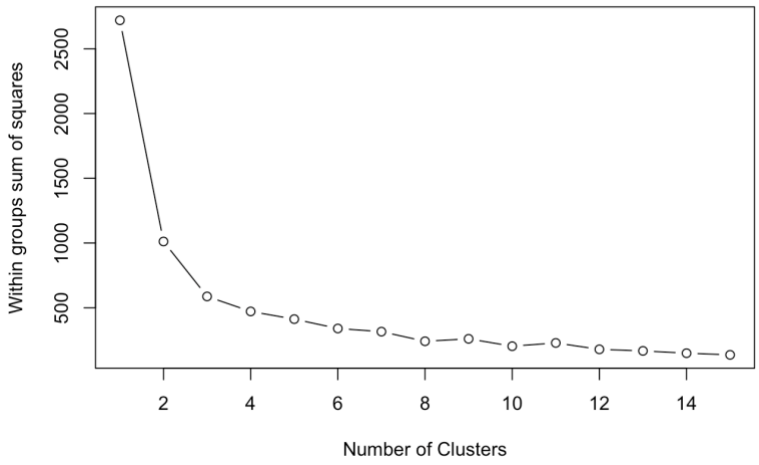

Figure 4. Graph of the squared distances from each data point to mean of cluster for each number of clusters.

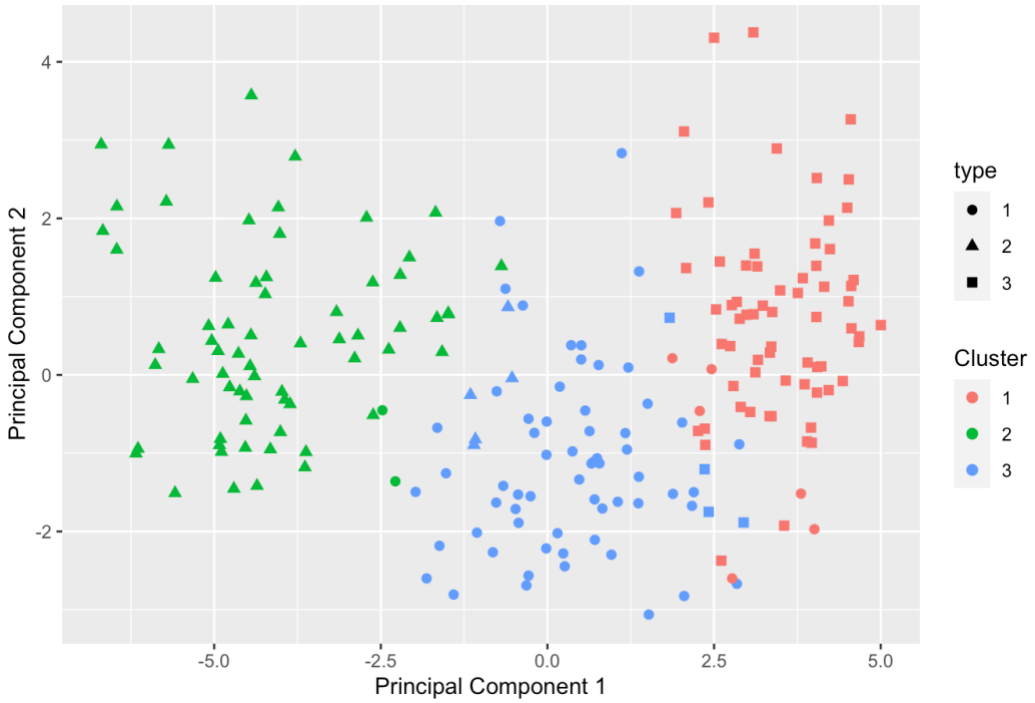

Figure 5. The results of k-means clustering with 3 clusters. The color represents the cluster assigned by k-means, and the shape represents the type of seed.

From fig 5, the k-means clustering formed the data into three clusters that do not overlap much. These three clusters, represented by color, closely resemble the true seed type for each data point, which is represented by the shape of the point. The error rate for the K-means clustering compared to the true type of seed was 8.05%, which

means that a data point was assigned to a cluster that does not match its true class 8.00% of the time, whereas it assigns it to the correct class 91.95% of the time.

C. Classification

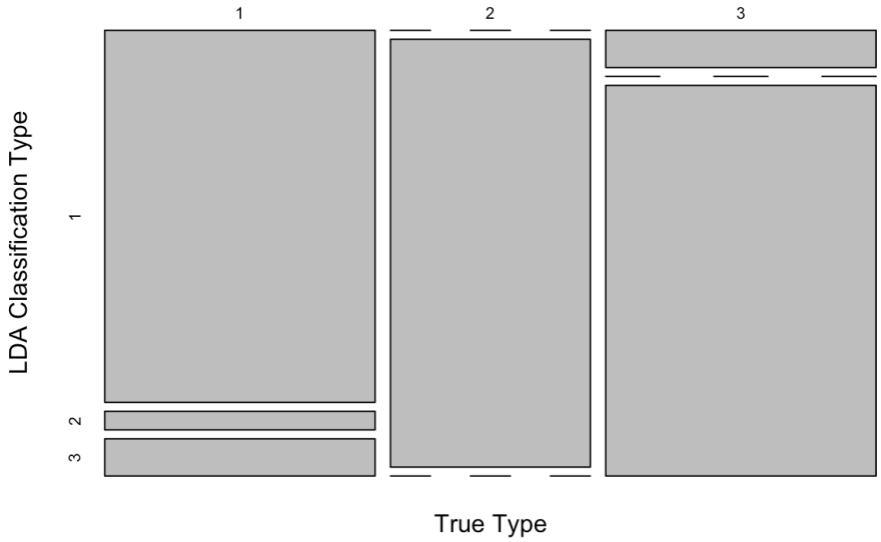

Figure 6. A Confusion Matrix showing a visual representation of the misclassification rate of Linear Discriminant Analysis. The y-axis is the class(type) assigned by LDA, and the x-axis is the true class(type).

Now I will perform several methods of classification on the dataset to see which performs the best at classifying the seed type on unseen data, or data in which we do not know the seed type of. For each method, I randomly sample the data into a training set with 70% of the data, and a test set with 30% of the data. I then train my model based on the training data, and then use it to predict the seed type of the testing data. Then I compare the seed type that each method assigned to the testing data to the actual seed types of the testing data. I then calculate the error rates for the methods, saving them into a data frame. I repeat this process 100 times to try to get closer to the true accuracy of the classification method and take the average error rate of each of the methods over the 100 repetitions of the classification.

The first classification method I used was Linear Discriminant Analysis. This resulted in a misclassification rate of 3.832%, and an accuracy of 96.181%.

The second classification method I used was multinomial logistic regression, which resulted in a misclassification rate of 4.8247%, and an accuracy of 95.1752%,

The third classification method I used was Kth nearest neighbors, which finds the k closest data points to a point and assigns it to whatever a majority of the k closest neighbors are assigned to. For this method, I had to pick the parameter k, which is the number of neighbors considered. I picked 5, 10, and 20, and classified the data with all 3 values 100 times. The results were:

1) K = 5 resulted in a misclassification rate of 1.4459%, and an accuracy of 98.554%,

2) K = 10 resulted in a misclassification rate of 2.436% and an accuracy of 97.56%,

3) K.= 20 resulted in a misclassification rate of 3.819% and an accuracy of 96.181%.

III. Analyzing the Automobile Data set

A. Exploratory Data Analysis

To begin analyzing the automobile data set, I will first plot a correlogram to get a general idea of the correlation between the variables in the data set. I am mainly interested in miles per gallon and price.

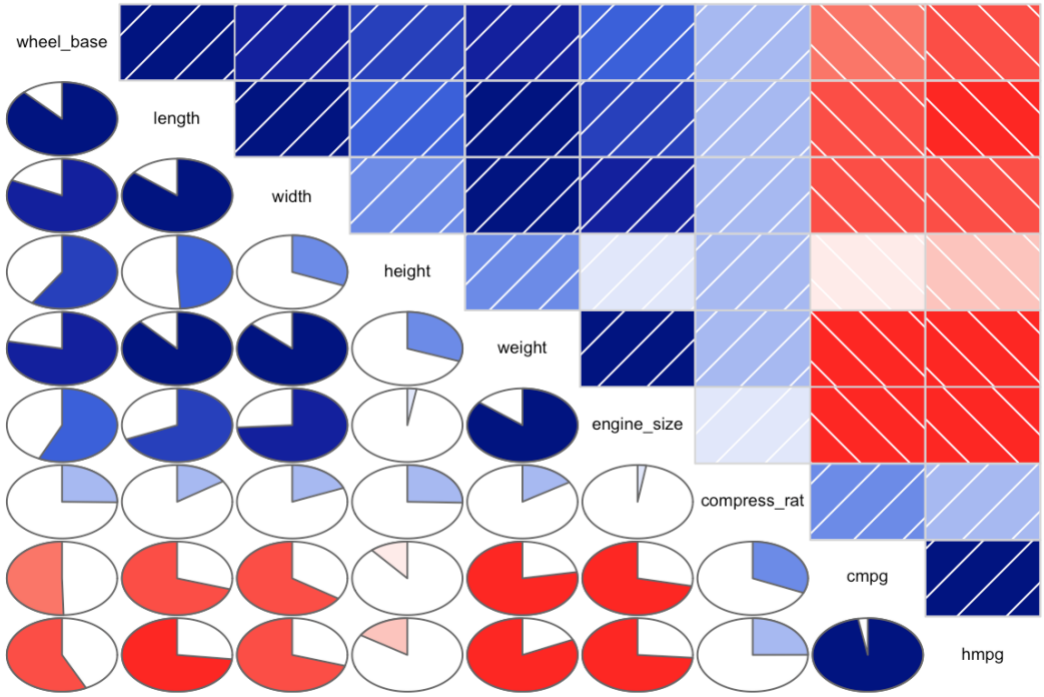

Figure 7. A correlogram of 9 variables of the Automobile data set. Blue represents a positive correlation, and red represents a negative correlation. Darker is stronger correlation, whereas lighter is a weaker correlation.

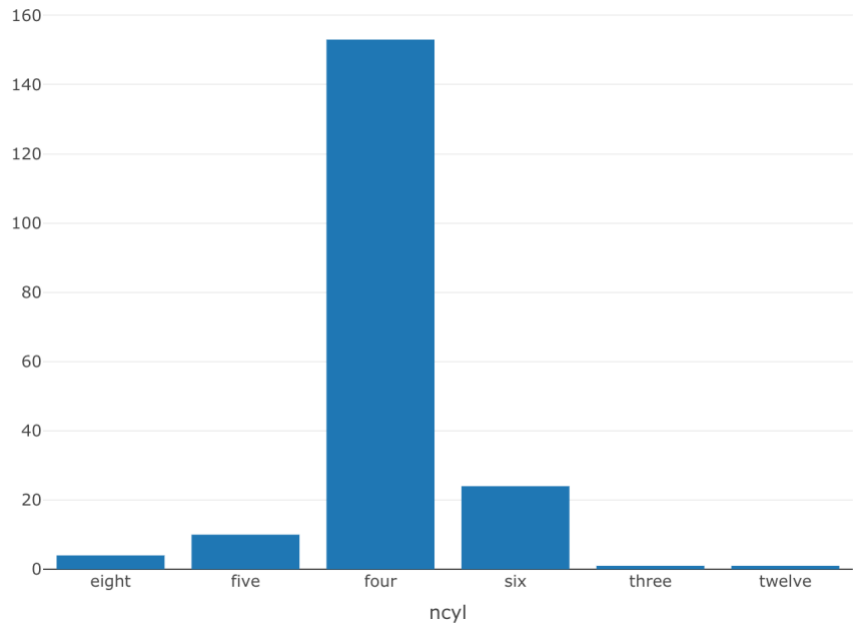

Figure 8. A histogram showing the number of cars with each number of cylinders.

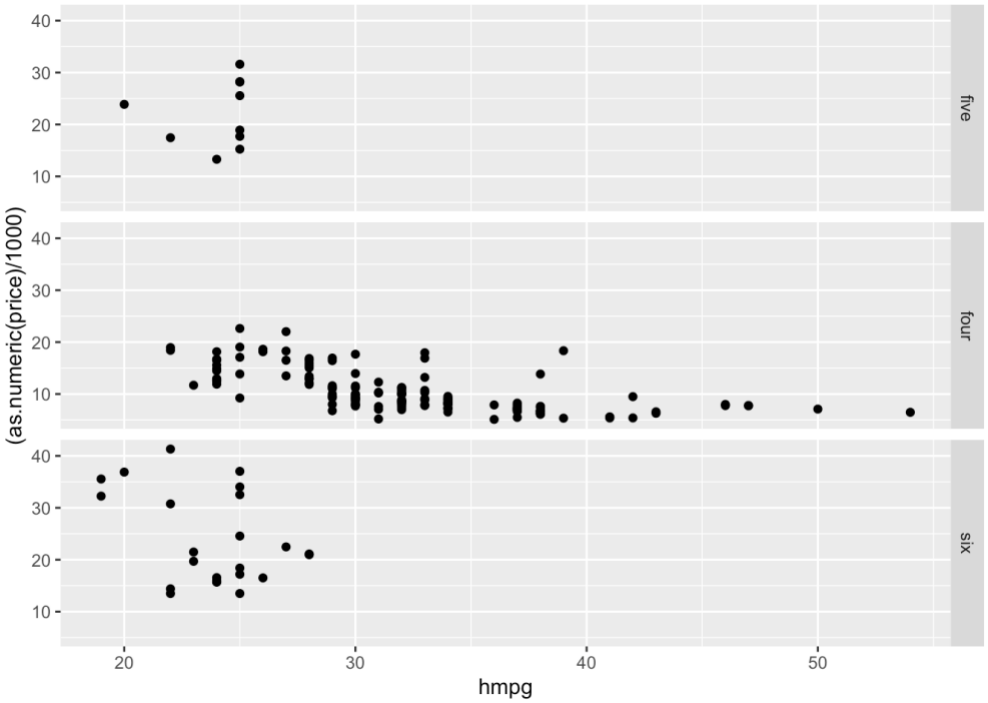

Figure 9. A scatterplot of highway miles per gallon versus price (in thousands), faceted on the number of cylinders.

Fig 7 shows that a lot of variables in the data set are highly correlated, such as engine size and weight. This can create problems such as multicollinearity, so I am going to be cautious about the regression coefficients of my prediction models.

Fig 8 shows that there are six factor levels for the number of cylinders, but three of them only contain 4 or less cars. This can create issues when performing prediction regardless of which method I use, because I will split the data into training and testing data, and since there is only one car with three cylinders, the training and test data will not have the same factor levels, creating issues. Because of this, I opted to remove factor levels which have less than 5 occurrences from the factor variables.

One of the primary objectives of this project is to analyze the effects of each variable on miles per gallons and predict the miles per gallon of a car. In the data set however, there is both city miles per gallon, and highway miles per gallon. Fig 10 shows the density plots of both city mpg and highway mpg. Highway miles per gallon and city miles per gallon have a correlation of 0.9719. This means they are very correlated, so I chose to focus on maximizing just one, highway miles per gallon.

From fig 9, you can see that as the price of a car increases, it's miles per gallon tends to decrease. Also, the only cars which have over 30 miles per gallon are ones with four cylinders, and the only cars with under 20 miles per gallon have six cylinders.

From fig 11, it is clear that the horsepower of a car has a strong negative relationship with the highway miles per gallon of a car. I will further explore and estimate the regression coefficients of variables on highway miles per gallon, .

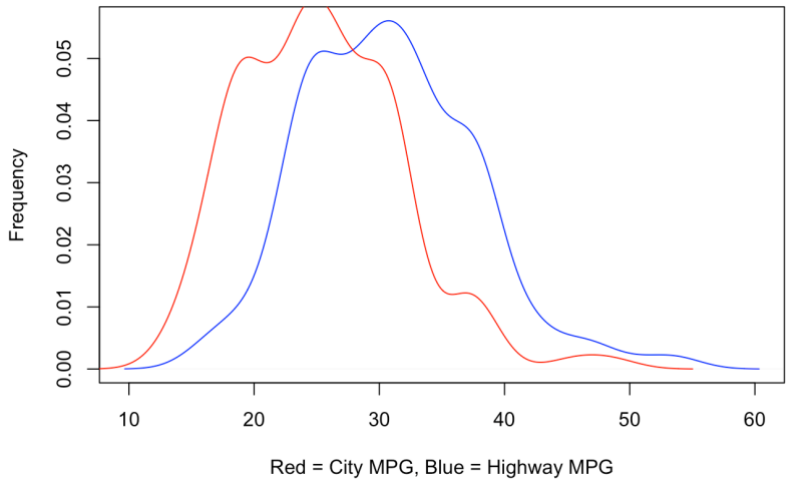

Figure 10. Density plot of the frequencies of both highway and city miles per gallon.

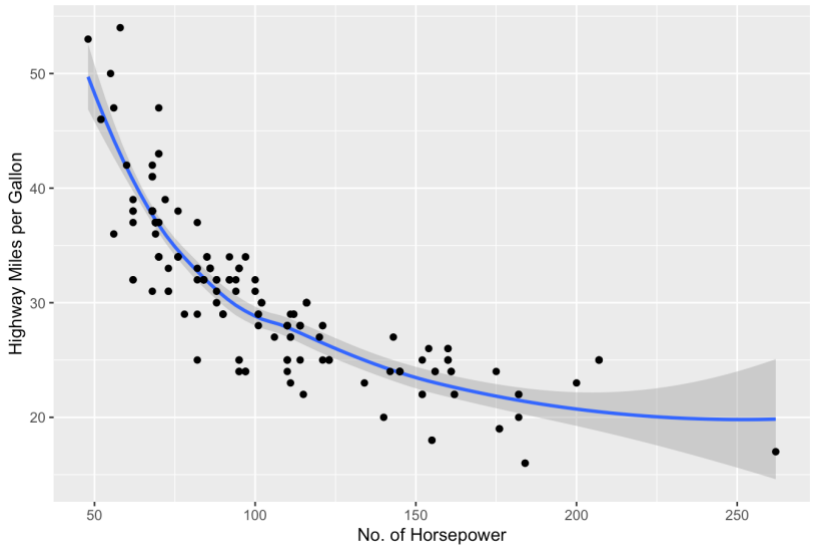

Figure 11. Scatter plot of highway miles per gallon versus number of horsepower, with a smoothed best fit line and a confidence band.

From fig 12 you can see a strong positive relationship between price and engine size. A strong correlation exists between many of the variables, which can create multicollinearity problems that disrupt the regression coefficients of the prediction models that involve multiple of these correlated variables.

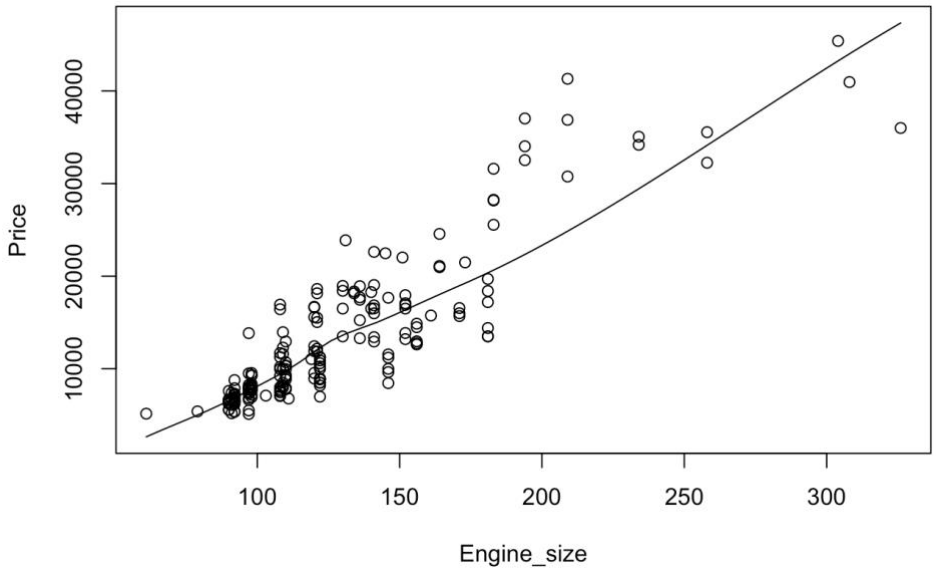

Figure 12. A scatterplot of price versus engine size with a best fit line.

```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5    PC6     PC7
## Standard deviation     2.7156 1.5341 1.13193 0.95375 0.78076 0.6125 0.60779
## Proportion of Variance 0.5268 0.1681 0.09152 0.06497 0.04354 0.0268 0.02639
## Cumulative Proportion  0.5268 0.6948 0.78637 0.85134 0.89488 0.9217 0.94807
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.49480 0.36888 0.34751 0.30104 0.25596 0.21676 0.14930
## Proportion of Variance 0.01749 0.00972 0.00863 0.00647 0.00468 0.00336 0.00159
## Cumulative Proportion  0.96555 0.97527 0.98390 0.99037 0.99505 0.99841 1.00000
```

Figure 13. The importance of each component of Principal Component Analysis. The proportion of variance is the proportion of variance from the original data set is captured in each principal component.

B. Principal Component Analysis

I will now perform Principal component analysis on the automobiles data set. For this, I will only use the numerical variables such as city miles per gallon, horsepower, price, weight, maximum rpm. The reduced data set without the categorical variables is 14 dimensions, so it consists of 14 variables. The table displayed in fig 13 provides the numerical results from the principal component analysis. You can see that the first principal component captured 52.68% of the variance in the original data, and principal component one and two capture 69.48%. The biplot in fig 14 shows the amount of variance that each variable contributes to principal component 1 and 2. This means that the data set can be reduced from 14 dimensions down to two, and still retain roughly 70% of the information contained in the non-reduced data set.

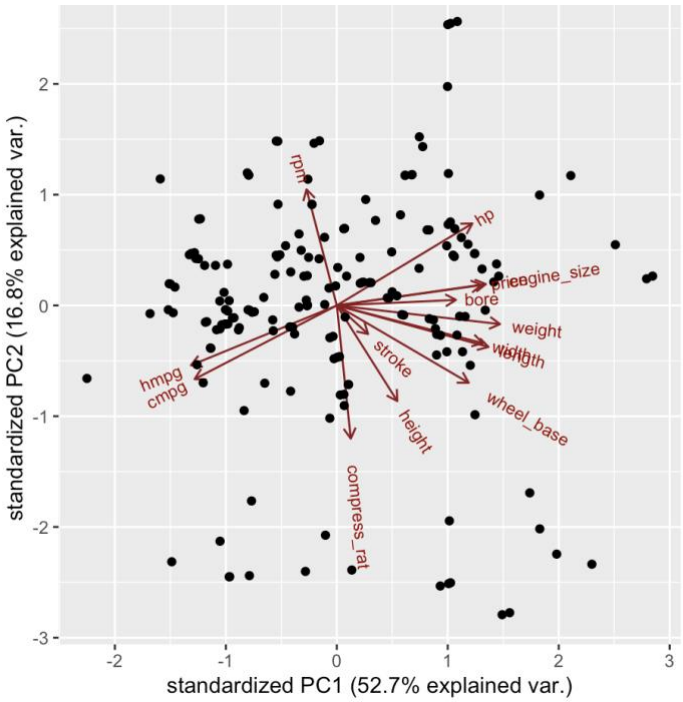

Figure 14. A biplot with the first and second principal components as the axes.

C. Prediction of Price

Next I aim to predict the price of a vehicle based on its other features. I begin by formatting the data to make it more suitable for prediction. I remove the factor levels of variables with less than 5 occurrences: cars with engine type rotor, cars with three, eight, or twelve cylinders, and cars with either spfi or mfi fuel systems. I also remove the variable engine location because almost all of the cars have a front engine. The reason I do this is because I am going to randomly sample training and testing data from the original data, and if one of them gets a factor level the other does not contain, the prediction functions will give an error or incorrectly predict some of the vehicles.

For all of the prediction methods used, I used the same process. First, I sample 70% of the data to be the training data, and the remaining 30% to be the testing data. I fit the model with prices as the response variable and every other variable as an explanatory variable to the training data, and then predict the prices of the testing data. I then compute the RMSE (Root Mean Squared Error) of each of the methods by taking the predicted price and subtracting the true price, dividing by the number of observations, squaring it, and adding up these values for all of the predictions. I repeat that process 100 times and take the mean error rate for each method over the 100 trials as the true error rate.

For attempt 1, I used linear discriminant analysis with price as the predictor and every other feature as explanatory variable in the model. I then performed ANOVA (Analysis of Variance) on the model, and removed fuel system, bore, compression ratio, and city miles per gallon as they had p-value less than 0.05, so they were not significant. I then performed linear discriminant analysis on the new model, which is attempt 2. For attempt 3 I performed linear discriminant analysis on a reduced data set without any of the categorical variables, which left me with 12 explanatory variables. For attempt 4, I performed ANOVA on the model without any categorical variables, and

found that bore, city miles per gallon, and highway miles per gallon are insignificant (p-value less than .05), so I removed them and predicted the prices again with linear discriminant analysis. For attempt 5, I predict the price using the random forest method of prediction. I used all of the variables available in the data set, including the categorical variables.

The RMSE (Root Mean Squared Error) represents the standard deviation of the residuals and is a good metric for seeing how accurate the predictions are by finding out how far the predicted values are from the true values, so I use this to compare the models. The results are:

1) Attempt 1 had a RMSE of 2903.0,

2) Attempt 2 had a RMSE of 2855.9,

3) Attempt 3 had a RMSE of 3151.68

4) Attempt 4 had a RMSE of 3124.886

5) Attempt 5 had a RMSE of 2202.50

So, the linear discriminant analysis with all of the variables outperformed the model without the categorical variables both times. For both the reduced and the full models for linear discriminant analysis, the RMSE slightly improved after removing some insignificant variables. The random forest method of prediction outperformed both of the linear discriminant models by a large margin. Using the weights from the random forest model, I can see that the three most important predictors in the price of a car are the make of the car, the engine size, and the weight. The three least important predictors are the number of doors, the fuel type, and the engine location.

D. Prediction of Highway Miles Per Gallon

Next I aim to predict the highway miles per gallon of a vehicle. I will use the same methodology as I used for price: 100 trials, random 70/30 training and testing data split, and finally take the average RMSE of each method after predicting the highway miles per gallon 100 times. For attempt 1, I will use linear discriminant analysis with highway miles per gallon as the response variable, and all of the other features as explanatory variables. For attempt 2, I perform analysis of variance on the previous model and remove insignificant variables (p-value less than 0.05): engine location, height, engine type, stroke, price, and predict the highway miles per gallon again using linear discriminant analysis. For attempt 3, I use the random forest method of predicting the highway miles per gallon of a vehicle, using the full data set.

The root mean squared error (RMSE) of each of the attempts are:

1) Attempt 1 had a RMSE of 1.4789,

2) Attempt 2 had a RMSE of 1.4770,

3) Attempt 3 had a RMSE of 1.79124

Both predictions using linear discriminant analysis outperformed the random forest prediction by a large amount. The prediction using linear discriminant analysis improved slightly after removing insignificant variables. The most important variables in predicting the highway miles per gallon of a car are city miles per gallon, horsepower, weight, and price. The least important variables in predicting the highway miles per gallon of a car are aspiration and number of doors.

III. Conclusions

Analyzing the Seeds data set was simple because it had a clear objective: classify the seed type. This was simple because all of the explanatory variables were numerical. Classification on the seed type was accurate because, as was shown in the exploratory data analysis, the seed types had relatively distinct clusters with little overlap. The seeds data set could be reduced to be less dimensions without losing too much variance because a lot of the variables are correlated. Randomly clustering the data with K means clustering almost perfectly assigned them to clusters that represented their seed type. The results from the Kth nearest neighbor classification method would have been more interesting if the data set was larger so that larger number of clusters would perform better.

Analyzing the automobile dataset was significantly harder than the seeds data set. It has 25 columns, contains a lot of missing data, and contains a mix of categorical and numerical data. It also contains many variables that are highly correlated, which can cause multicollinearity issues, distorting the regression coefficients of the models. When predicting the price of a car using linear discriminant analysis, I performed it both with and without the categorical variables, but the model without them performed worse, so I had to work with them. I opted to ignore the symbolling variable as it was a pre-computed class, and I do not consider it to be a true label such as seed type. I removed the variable normalized losses before performing any data analysis because it contained a large number of NA's, and because the data set was already small with 205 vehicles, I did not want to reduce it further.

References

[1] J. Aitchison, I.R. Dunsmore, "Statistical Prediction Analysis," Cambridge University Press, 1975.
[2] C. Maklin, *Random Forest in R*, Towards Data Science, Jul 2019
[3] https://archive.ics.uci.edu/ml/datasets/Automobile
[4] https://archive.ics.uci.edu/ml/datasets/seeds
[5] F. Hsieh, Data Science Practice, unpublished. Department of Statistics, University of California, Davis, April 2020.