STA160 Midterm Report

Zeng Fung Liew (913802324)
lzfliew@ucdavis.edu

April 26, 2020

Abstract

The seeds dataset describes various characteristics of wheat kernels of the Kama, Rosa and Canadian variety extracted via X-ray imaging. The goal of this analysis is to determine classes of each kernel based on its characteristics. Data visualizations of each characteristic is portrayed in the form of gapped histograms with variable bin widths before conducting multivariate analysis by implementing linear discriminant analysis (LDA) on the first two principal components. The results of the classifications are then further analyzed. The automobile dataset describes the specifications, price and risk rating of imported cars in 1985. The goal of this analysis is to determine the main attributes that contributes to the price of cars. Visualizations such as histograms and scatterplots are used for the analysis and a model is fitted using step-wise regression technique. The importance of each attribute in determining the price of a car is then discussed.

1 Seeds Dataset

In this dataset, we examine three different varieties of wheat kernels Kama, Rosa and Canadian. 70 kernels of each wheat variety were randomly sampled and analyzed based on the various attributes which were obtained via the visualization of internal kernel structures using the soft X-ray technique.

1.1 Methods

The analysis of this dataset will focus mainly on the effectiveness of single variate and multivariate classifications.

Single Variate Analysis

For each variable or attributes in the dataset, an empirical cumulative distribution function (ECDF), a gapped histogram, and a DESS function will be plotted. The goal of this is to display a histogram with the best binwidth. A histogram with a binwidth too narrow or too wide fails to portray the data optimally. Hence, a DESS function is used to determine the binwidth of the gapped histogram in order to help us analyze and classify the data. Several red lines are then drawn on the ECDF to indicate the bins of the histogram.

Multivariate Analysis

Since we know that the classification of wheat kernel varieties is not based off a single attribute, a multivariate analysis has to be conducted to analyze the interaction of the attributes in determining the varieties of wheat kernels.

However, we know that it is difficult to visualize the interaction of data since it has such a large number of attributes. To counter this, a principal componenet analysis (PCA) is conducted to reduce the current dataset into a 2-dimension data for easier visualization of data classification. Then, we will conduct linear

discriminant analysis (LDA) and determine the apparent and expected actual error rates based on its analysis.

Lastly, we want to figure out if the three wheat kernel varieties can be classified into other different types of categories. To do so, we will conduct hierarchical clustering and analyze the number of clusterings.

1.2 Results and Analysis

Univariate Analysis: ECDF, Histograms, and DESS

Figures 1, 2, 3, 4, 5, 6, and 7 show the empirical CDF, the gapped histogram and the DESS plot of each attributes. For histograms, note the following color codes:

Red	Kama
Green	Rosa
Blue	Canadian

Table 1: Wheat varieties and their color codes

Additionally, in the DESS plots, the red lines describe the threshold $\frac{(b_j - a_j)^2}{3}$ whereas the blue lines describe the DESS for each bin of the histograms. Based on figures 1 to 7, we can see certain attributes

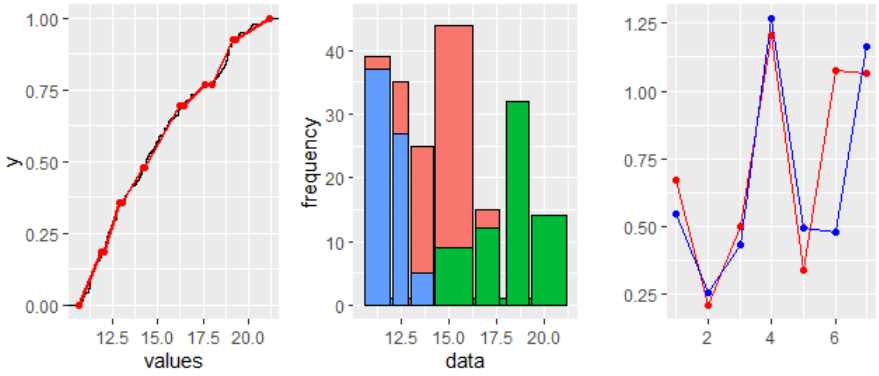


Figure 1: Empirical CDF, Gapped Histogram and DESS plot for area of kernel

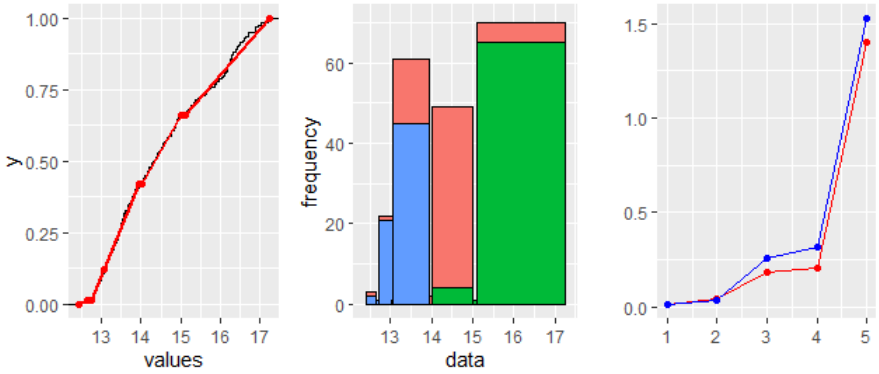


Figure 2: Empirical CDF, Gapped Histogram and DESS plot for perimeter of kernel

such as kernel area, perimeter, length and width, and the length of kernel grooves can be good predictors

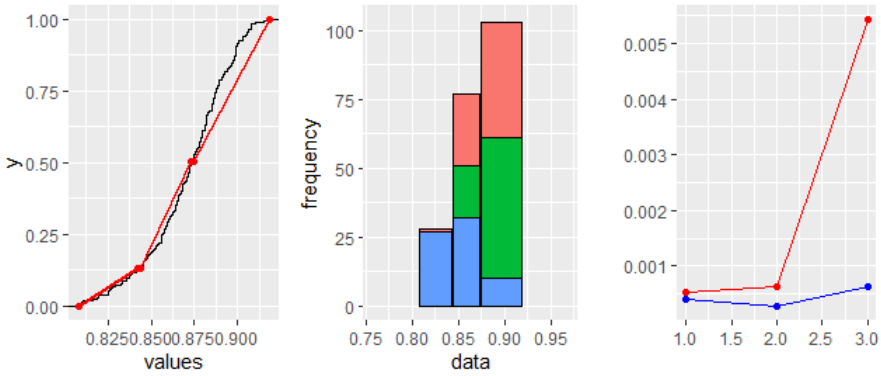


Figure 3: Empirical CDF, Gapped Histogram and DESS plot for compactness of kernel

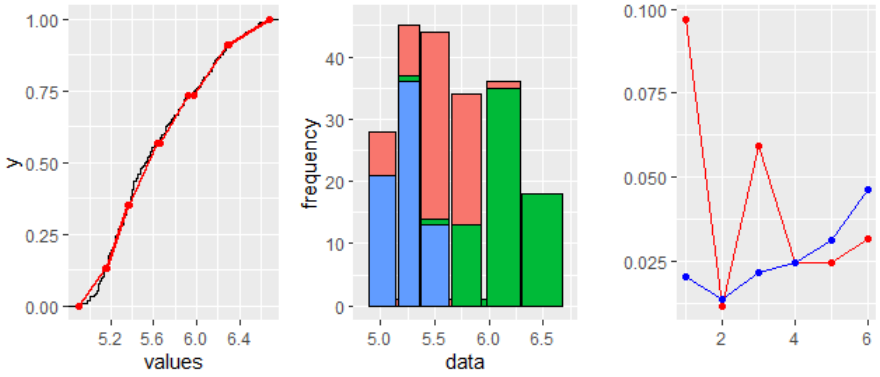


Figure 4: Empirical CDF, Gapped Histogram and DESS plot for length of kernel

to classify the Rosa variety. In these histograms, we can see that most of the right bins are predominantly occupied by the Rosa data points. The same cannot be confirmed about the other two varieties since the other bins in the some of those histograms are filled evenly with Kama and Canadian data points. However, we can infer that the size of the Kama wheat kernel is between the kernel sizes of Rosa and Canadian wheat since the middle intervals of Figures 1, 2, and 4 contain a larger proportion of Kama variety. In short, the kernel size of the Rosa wheat is significantly larger that the other two varieties. As for the compactness of kernels in Figure 3, while Rosa kernels are still considered highly compact, there is no significant difference when compared to the other two varieties. Similarly, while the higher values of the assymetry coefficient of the kernels in Figure 6 mainly consist of the Canadian variety, there is no significant difference when compared to the other two varieties since all varieties are distributed quite evenly at bins between 2 and 4.

1.2.1 Multivariate Analysis: PCA and LDA

We conduct principal component analysis (PCA) on the standardized data and analyze the loadings of the first two principal components in Table 2 to make inferences of the importance of each characteristic.

Table 2: Loadings of the first two principal components

Comp.	Area	Perimeter	Compactness	K. Length	K. Width	Assym. Coef	K. Groove Length
1	0.444	0.442	0.277	0.424	0.433	-0.199	0.387
2	0	0	-0.529	0.206	-0.117	0.717	0.377

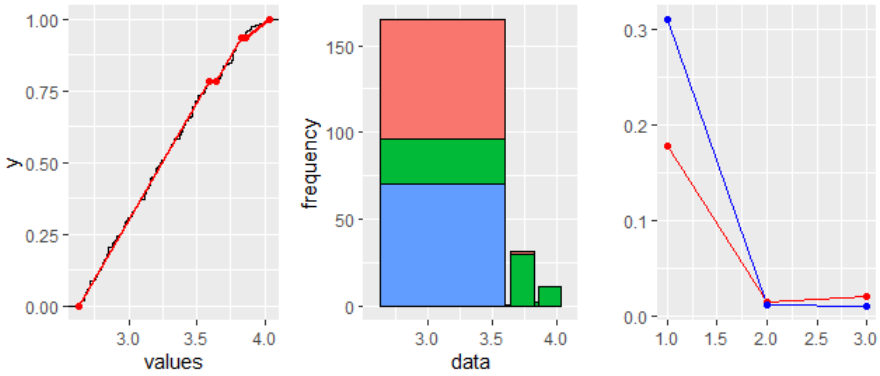


Figure 5: Empirical CDF, Gapped Histogram and DESS plot for width of kernel

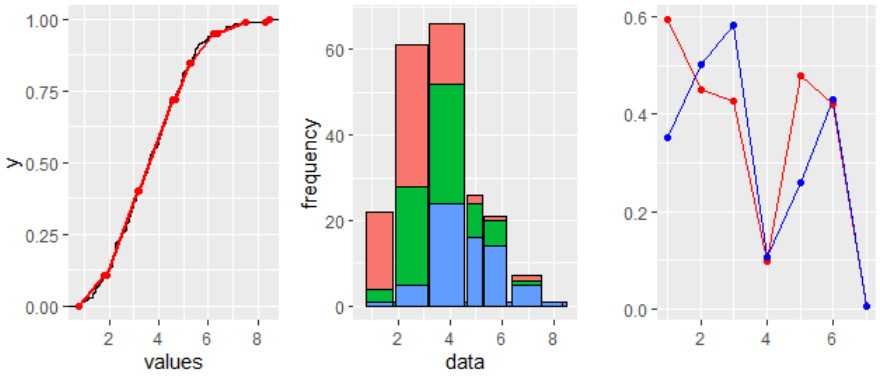


Figure 6: Empirical CDF, Gapped Histogram and DESS plot for assymmetry coefficient of kernel

The first principal component contains larger values close to 0.4 in characteristics such as kernel area, perimeter, length, width, and groove length. What this means is that the average of this attributes is the most significant in determining the variety of wheat kernels. On the other hand, the second principal component consists mainly of the compactness (≈ 0.5) and the assymmetry coefficient (≈ 0.7) of the kernel. In other words, the physical size of wheat kernels is what separates the different varieties the most, as inferred in the previous section.

Additionally,we find that the proportion of the total variance due to the first two principal components is close to 89%. This means that we can replace all the attributes of the data with the first two principal components without much loss of information. In doing so, it is easier for us to visualize the data since we can plot the data in two dimensions, as shown in Figure 8a.

It is clear that there are distinct differences between all three wheat varieties since it is obvious that the data is spread into three main clusters. We then conduct linear discriminant analysis (LDA) to determine its accuracy of our classifications.

Upon conducting LDA, we construct a confusion matrix as shown in Table 3 to determine the error rate.

Table 3: Confusion matrix from LDA

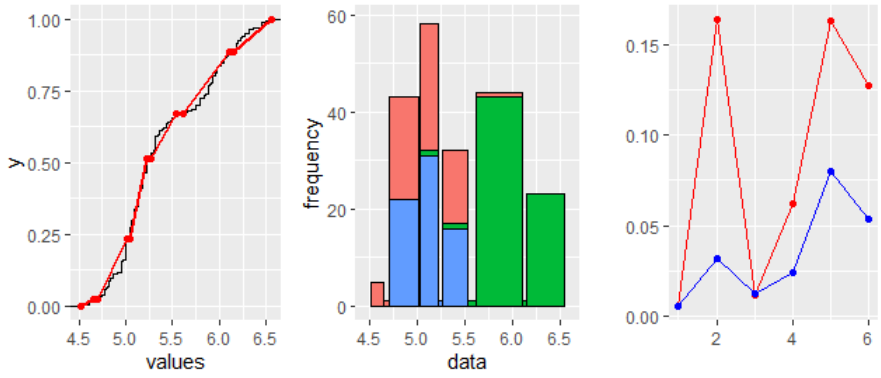
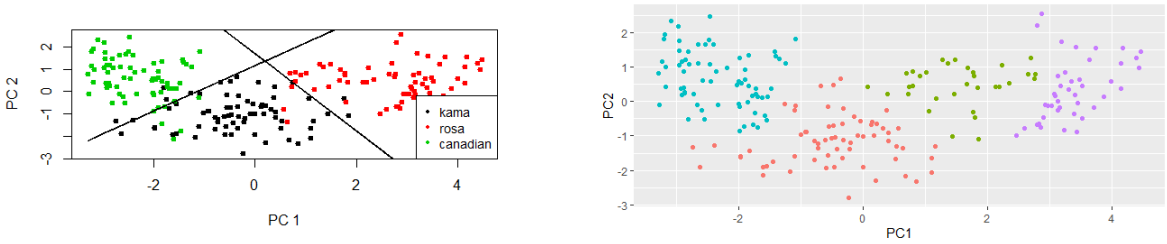


Figure 7: Empirical CDF, Gapped Histogram and DESS plot for length of kernel groove



(a) Classification of kernels based on the first two principal components

(b) Hierarchical clustering of the first two principal components using average linkage

Figure 8: Classification and clustering on first two PC

Actual \ Predicted	Kama	Rosa	Canadian
Kama	62	3	5
Rosa	3	67	0
Canadian	5	0	65

Apparent Error Rate $= (\frac{3+5+3+5}{210}) \times 100\% = 7.619\%$

We then go a step further and perform LDA via Lachenbruch's holdout to determine an estimated actual error rate. In Lachenbruch's holdout, we first leave out the first row of our data (call it our test data) and perform LDA on the rest of the dataset. Based on the boundaries we obtain in that LDA, we use them to classify our test data. We then repeat these steps for the rest of the dataset. Similar to conducting LDA previously, we then construct a confusion matrix to determine our actual error rate. The idea of Lachenbruch's holdout is to find an estimate for our classification actual error rate.

Table 4: Confusion matrix from LDA

Actual \ Predicted	Kama	Rosa	Canadian
Kama	62	3	5
Rosa	3	67	0
Canadian	6	0	65

Estimated Actual Error Rate (EAER) $= (\frac{3+5+3+6}{210}) \times 100\% = 8.095\%$

Based on tables 3 and 4, we see that the error rates are relatively low (less than 10%), this suggests that the first two principal components is sufficient in classifying the wheat varieties. It also suggests that the

differences between wheat varieties are quite distinct. Additionally, the apparent error rate is also close to the estimated actual error rate since the sample size in this data is large.

1.2.2 Multivariate Analysis: Clustering

Once again, using the first two principal components, we perform hierarchical clustering with average linkage. In the end, we find that the data can be distinguished into four clusters, as shown in figure 8b. However, we can see that figure 8b is similar to figure 8a, except that the Rosa data points has been split into two colors. This suggest that there could be two sub-varieties within the Rosa wheat variety.

1.3 Conclusion

In this dataset, both univariate and multivariate analysis have provided us with similar results. We find that the size of kernels (ie. kernel area, perimeter, width, length and length of kernel groove) are the most significant determinants in classifying the three varieties of wheat. This inference based on histograms is then strengthen by the results of principal component analysis. While we may find that the size of kernels increases in the order of Canadian, Kama and Rosa, it is hard to determine the boundary in classifying Canadian and Kama wheats based on kernel size as compared to Rosa and both Kama and Canadian. However, it is also important to note that a kernel's compactness and asymmetry coefficient are also important factors in the classification of wheat, as shown in the second principal component in Table 2.

When we perform linear discriminant analysis on the first two principal components, we find that the accuracy of the prediction of wheat variety is still relatively high at about 90%. This is largely due to what the first two principal components represent, which is 89% of total variance of the data.

Lastly, based on the hierarchical clustering performed on the first two principal analysis, we find that there is a total of four distinct clusters, of which two of them were within the Rosa data points. This suggests that there is a possibility that there exists two sub-varieties within the Rosa wheat variety.

2 Automobile Dataset

The Automobile data contains various details of imported cars in 1985, which includes specifications, prices, risk rating, and normalized loss. In this analysis, we will focus on how car specifications differ based on its respective make and price.

2.1 Methods

The analysis of this dataset is largely dependent on `ggplot` visualizations in R. Due to the large number of car make in this dataset, visualizations based on each category can be very complicated. To tackle this problem, we classify each car make into three tiers based on its price range, with Tier 1 containing car makes with the highest prices and Tier 3 with the lowest.

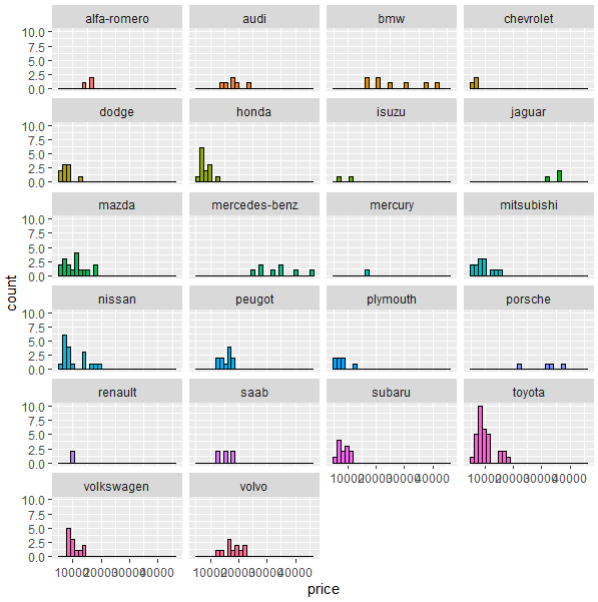


Figure 9: Price histograms by car make

Then, we produce data visualizations on each specification and analyze how it differs by tiers. For categorical attributes, we construct three rows of histograms of price separated by tiers, with the attributes being the color visualizations. We then try to search for existing trends within those histograms, such as finding certain attributes more prevalent in higher price ranges or higher tier cars. For continuous attributes, we construct colored scatterplots between the attributes and price to determine the their relationships.

After we conduct our data visualizations, we fit a linear model onto the dataset using stepwise regression to determine the best subset of attributes to determine the price of cars. We then use it compare with our previous analysis before making a conclusion.

2.2 Analysis

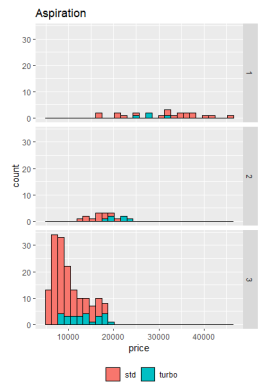
2.2.1 Splitting into tiers

In Figure 9, we constructed histograms of price by grids of car make. Each of those grids are scaled equally. A vehicle brand that produces cheaper cars have histograms that are skewed to the left, and vehicle brands that produce more expensive cars have histograms that are skewed to the right. Based on the price range of each car make, we categorize the car make into three tiers in Table 5.

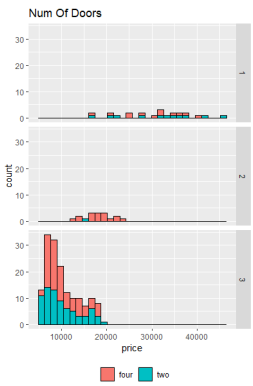
Table 5: Tiers of cars with respective car makes

Tier	Make
1	BMW, Jaguar, Mercedes-Benz, Porsche
2	Audi, Volvo
3	Alfa-Romero, Chevrolet, Dodge, Honda, Isuzu, Mazda, Mercury, Mitsubishi, Nissan, Peugot, Plymouth, Renault, Saab, Subaru, Toyota, Volkswagen

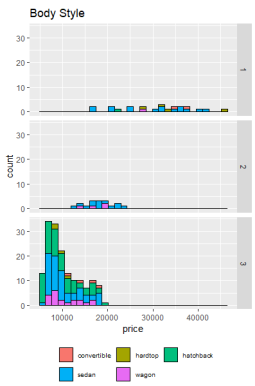
2.2.2 Analysis of relationship between car price and categorical variables



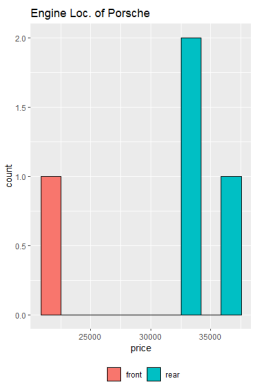
(a) Price histogram by aspiration



(b) Price histogram by number of doors

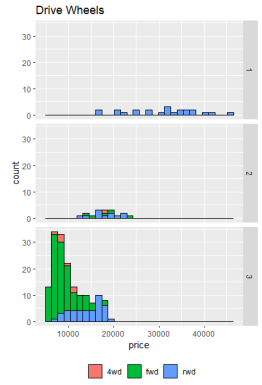


(c) Price histogram by body style

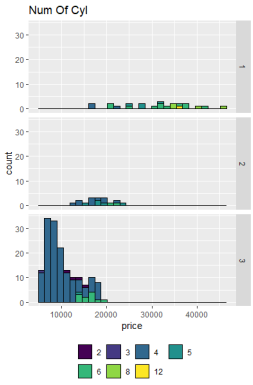


(d) Bar chart of engine location in Tier 1 cars

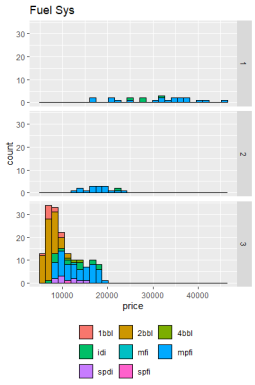
Figure 10: Analysis of relationship between price and categorical variables (Part 1)



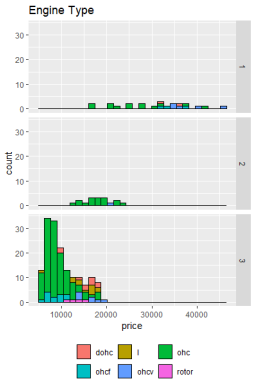
(a) Price histogram by drive wheels



(b) Price histogram by number of cylinder



(c) Price histogram by fuel system



(d) Price histogram by engine type

Figure 11: Analysis of relationship between price and categorical variables (Part 2)

Moving on, we analyze each car attribute within tiers and determine its relationship with its price in Figures 10 and 11. In Figure 10a, we notice that cars with Turbo aspirations tend to appear in higher proportions in higher price ranges among cars of Tiers 2 and 3. However, this does not seem to be the case among Tier 1 cars, where the price range of Turbo cars lies in the middle of the price range of Standard cars. This could be due to other attributes that result in such high prices among Standard cars in Tier 1. As for Figure 10b and 10c, the distribution of car prices with two and four doors and different body styles seems quite even, suggesting that neither the number of doors nor the body style of a car has a significant impact on its price. We then look into the engine locations of cars, it is interesting to find that only Porsche produces cars with rear engines. While this might suggest that engine location is not a determinant of price, we find that Porsche cars with rear engines are at least $10,000 more expensive than the Porsche car with front engine, as shown in Figure 10d.

As for Figure 11, we can see that there exists some trends within drive wheels, number of cylinders, fuel system, and engine type. In Figure 11a, we can see that the proportion of forward-wheel drive cars is larger

towards the left of the histograms, and the proportion of rear-wheel drive cars increases towards the right of the histograms. Also, it is interesting to note that Tier 1 vehicle brands only manufacture rear-wheel drive cars, and that Tier 3 brands manufacture majority of the forward-wheel drive cars. Additionally, in the price histogram in Tier 3, the proportion of forward-wheel drive cars are skewed to the left and rear-wheel drive cars are skewed to the right. A similar statement can be made about the number of cylinders in a car (Figure 11b), where majority of the cars contain four cylinders, and the cars that contain a higher number of cylinders are mostly in Tiers 1 and 2, or occupy a higher price range in Tier 3. As for fuel systems (Figure 11c), we see that the majority of the cars use either the 2bbl or the mpfi fuel system. The difference that we observe between them is that the mpfi system is usually used in more expensive cars, especially those of Tiers 1 and 2, whereas the 2bbl system is used mostly in cars cheaper than $15000. Lastly, the most used engine type by cars is 'ohc', as shown in Figure 11d. While it has a large price range on all three tiers, a large proportion of cars made with 'ohc' engine tend to be cheaper. On the other hand, cars with 'ohcv' engines tend to be more expensive as they tend to exist near the right side of the histogram distributions. This is more prevalent in Tiers 1 and 3.

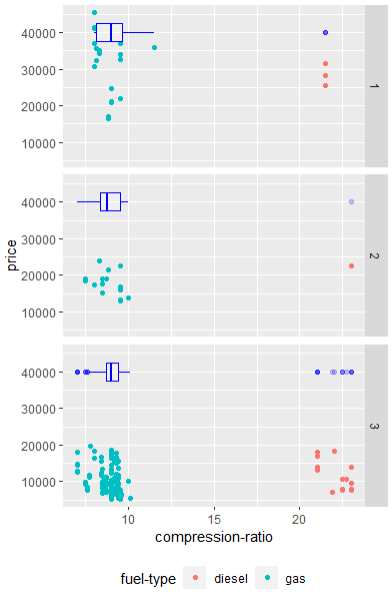


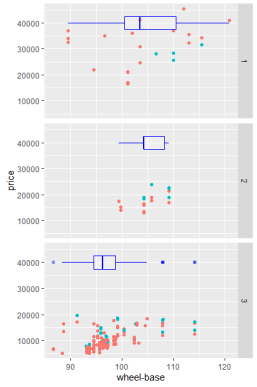
Figure 12: Compression ratio vs price scatterplot

In short, we say that drive wheels, number of cylinders, engine type and fuel systems are the main categorical contributers to the price of a car. While the aspiration of a car might suggest an impact on its price, it is only applicable to cars in Tiers 2 and 3. As for engine location, while it seems to be a significant factor for determining the price of a car, the sample size of cars with rear engines is far too small to make such a conclusion.

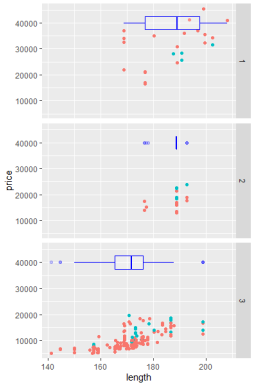
2.2.3 Analysis of relationship between car price and continuous variables

We then move on to analyzing the continuous attributes of the cars with histograms and boxplots as shown in Figure 13. Here we can see the distributions of each attribute, separated into 3 tiers. In most of these plots, we can see significant differences between tiers. Cars in Tiers 1 and 2 are longer,wider and heavier with larger wheel bases, engines, bores and horsepower as compared with Tier 3 cars. However, there is a drawback, as shown in the Figure 13k and 13l,where cars in Tiers 1 and 2 seem to be less fuel-efficient. Another interesting finding is the compression ratios of cars in Figure 12. Regardless of the price of cars, the distributions of compression-ratio is split very distinctly into two, with cars that use diesel having a compression ratio of less than 10 and a cars that use gas having a compression ratio larger than 20, with nothing in between.

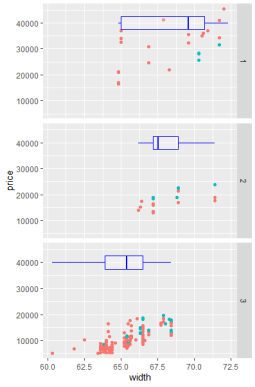
Based on how far the box-plots are from between tiers for each attribute, we conclude that the most influential continuous attributes in determining the price of a car are its curb weight, engine size, length, width and horsepower.



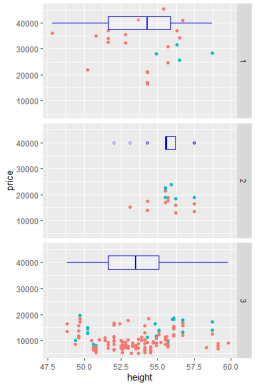
(a) Wheel base vs price scatterplot



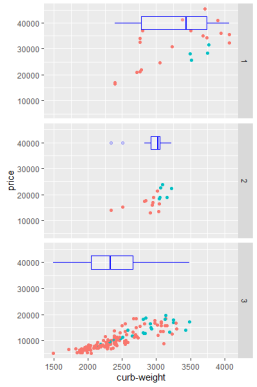
(b) Length vs price scatterplot



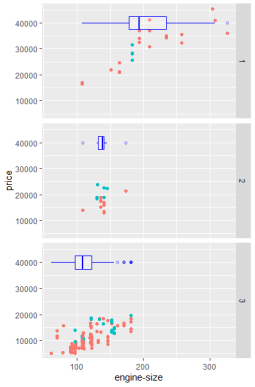
(c) Width vs price scatterplot



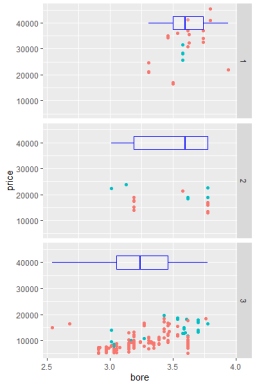
(d) Height vs price scatterplot



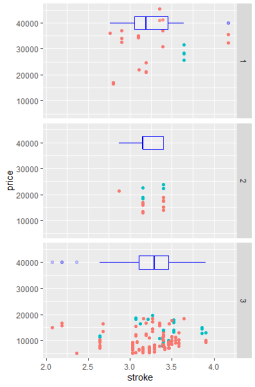
(e) Curb weight vs price scatterplot



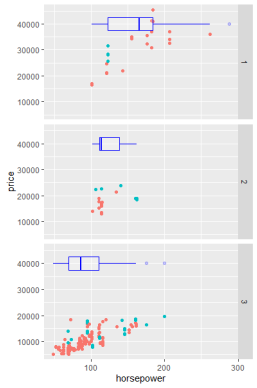
(f) Engine size vs price scatterplot



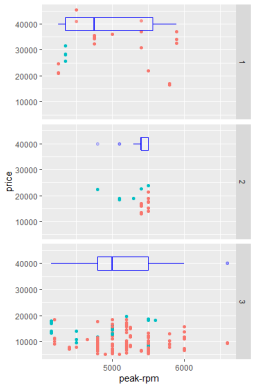
(g) Bore vs price scatterplot



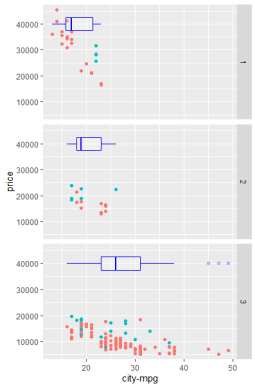
(h) Stroke vs price scatterplot



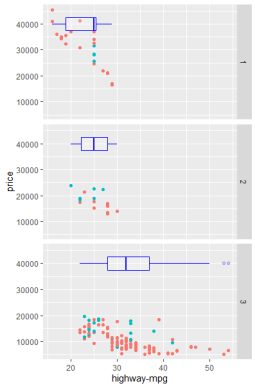
(i) Horsepower vs price scatterplot



(j) Peak rpm vs price scatterplot



(k) City mpg vs price scatterplot



(l) Highway mpg vs price scatterplot

Figure 13: Analysis of relationship between price and continuous variables by tiers and aspiration (Red data points represent Standard cars, Blue data points represent Turbo Cars)

2.2.4 Fitting a linear model

The next thing we want to do is to fit a linear model of the price of a car based on its attributes. We do so by implementing step-wise regression starting from the full model containing all attributes. We then drop or add attributes into the model based on the F-values of each attribute. The attribute with the largest F-value is dropped if it is larger than a threshold of $\alpha = 0.01$ and the attribute with the lowest F-value is added into the model if is smaller than the same threshold. This step is repeated until we arrive at our optimal model.

After running the step-wise regression, our best model consists of the following subset of attributes:

- make
- width
- engine type
- engine size
- aspiration
- curb weight
- number of cylinder
- peak rpm

With this model, we obtain an R^2 statistic of around 0.95. This means that about 95% of the total variation can be explained by the eight attributes above, which suggest that the model is a very good fit.

What is interesting about the above subset is that the attribute 'peak rpm' that is considered as one of the best predictors of the price of a car, as oppose to our conclusion in Figure 13j. The rest of the subset however, is more or less of what we expected.

2.3 Conclusion

Base on this analysis, we can conclude certain dominant attributes in cars that determine their prices. First off, the make of a car is a huge factor to the wide price ranges. Car make such as Mercedes-Benz, Porsche and Jaguar typically make cars that cost more than $30,000, whereas we can go to Toyota, Nissan or Mazda for cars that are half the price. As for aspiration and engine type, turbo cars with 'ohcv' engines are generally more expensive than standard cars with 'ohc' enginces.

Additionally, attributes such as drive wheels, number of cylinders, and fuel system in cars can be huge determinants of car prices. Rear-wheel drive cars with higher number of cylinders that utilize the mpfi fuel system typically costs more than forward-wheel drive cars with four cylinders and utilize the 2bbl fuel system.

However, there are certain attributes whose influence on price are up for debate. While the data visualizations in Figures 13a, 13g and 13i display certain influence on the price of car by wheel base, bore and horsepower, they are not considered in the subset of attributes for the best linear model. On the other hand, the peak rpm of a car, whose scatterplot with price in Figure 13j does not display much of a correlation, is included in the subset of attributes for the best linear model.

Lastly, while it is obvious that expensive cars are larger and more powerful than cheaper cars, they come at another expense of having lower fuel-efficiency, which can be a deal-breaker for people who travel a lot.

3 Reference

[1] ECDF, Gapped Histogram and DESS function codes by Tania Roy
http://anson.ucdavis.edu/~taniaroy/taniaroy/ANOHT_functions.R
[2] R-bloggers - Classification with Linear Discriminant Analysis
https://www.r-bloggers.com/classification-with-linear-discriminant-analysis/