Technical University of Denmark

02540: Introduction to Machine Learning and Data Mining
Project One:

Erik Gylling, s173896
Christian Hinge, s174173

March 5th, 2019

Contents

Introduction

In this report we will explore our chosen data set about cars and visualize its components in order to get a better understanding of the data and as a preparation for future analysis using various techniques from machine learning. The main objective right now is to extract features and discover hidden connections through visualization of the data.

Data set

Our data set consists of 205 cars (observations) and 26 different attributes such as brand, price, horsepower, and wheel-base. Some of the attributes contain continuous values others while others contain discrete integers or categories.

The data is retrieved from "http://archive.ics.uci.edu/ml/datasets/Automobile"
The creator is *Jeffrey C. Schlimmer*, who wanted to know if there were any connections between the car characteristics, the safety of the car and the decreasing value of the car per year. He has not published any results.

In 1989 *D.Kibler, D.W.Aha, & M.Albert* were working with this data set. They created an instance-based prediction of real-valued attributes. The purpose was to predict the price of car using all numeric and boolean attributes. Their method was to use an instance-based learning (IBL) algorithm derived from a localized k-nearest neighbor algorithm. Compared with a linear regression prediction so all instances with missing attribute values were discarded. This resulted with a training set of 159 instances, which was also used as a test set (minus the actual instance during testing).
Results: Percent Average Deviation Error of Prediction from Actual [1]

- 11.84 for the IBL algorithm
- 14.12 for the resulting linear regression equation

So far we have discussed using both the *price* and *make* attribute as the class labels. Upon analyzing the data however, it is clear that *make* contains more than 20 discrete values, some of which have a frequency of only one in the whole dataset. Therefore, just like the group in 1989, we will focus on the *price* attribute - possibly transforming it into discrete price categories.
The primary machine learning modelling aim is to create a classification for the price and to detect anomalies in this attribute. To fulfill this aim, we believe that attributes such as*price*, *cylinders*, *engine-size*, *curb-weight*, *city-mpg* and *body-style* are especially relevant, because pricey cars tend to use more fuel, be bigger in size, have more cylinders, have a special shape, and a bigger engine.

Attributes

In the section below is shown some attributes representative of the data set. The rest of the attributes are listed and described in the Appendix.

- **price** is a continuous ratio attribute with values from: 5118 to 45400.
  This attribute specifies the price of car. One may classify it as discrete since the price is specified in an integer amount. However, due to the enormous range, this attribute is also included when we refer to the "continuous attributes".

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names

- **num-of-cylinders** is a discrete ratio attribute with the values: eight, five, four, six, three, twelve, two.
  This attribute specifies how many cylinders the car enginge has.
- **engine-size** is a continuous ratio attribute with values from: 61 to 326.
  This attribute specifies the size of the engine.
- **curb-weight** is a continuous ratio attribute with values from: 1488 to 4066.
  This attribute specifies the curb-weigth of the car.
- **city-mpg** is a continuous ratio attribute with values from: 13 to 49.
  This attribute specifies how many miles the car drives per gallon of fuel in the city.
- **body-style** is a discrete nominal attribute, with the categories: hardtop, wagon, sedan, hatchback, convertible.
  This attribute specifies the styling of the car.

The attributes: *Make*, *Fuel-type*, *Standard Engines*, *Aspiration*, *Body-style*, *Drive-Wheels*, *Engine-location*, *Engine-type* and *Fuel-system* have been converted to One-out-of-K columns, which have placed in the back of the dataset. These attributes are all nominal, so the k-encoding is more sensible for both the PCA and for future data analysis. The values of *num-of-doors*, *symboling*, and *num-of-cylinders* have been converted from strings to integers.

Data issues

- 41 values are missing in the *normalized-losses* attribute.
- 2 values are missing in the *num-of-doors* attribute.
- 4 values are missing in the *bore* attribute.
- 4 values are missing in the *stroke* attribute.
- 2 values are missing in the *horsepower* attribute.
- 2 values are missing in the *peak-rpm* attribute.
- 4 values are missing in the *price* attribute. The missing values have been replaced with NaN.

For now, it has has been decided to remove the *normalized losses* attribute due to the large amount of NaN values. Furthermore, the 12 observations with NaN values have been removed, as we did not feel that we could justify replacing the values with means or medians. We may reverse some of these changes later if we obtain new knowledge. After 1-out-of-k-encoding and data removal, our data set contains 193 observations and 64 columns.

Basic Statistics

In the following table mean, variance, and standard deviation have been calculated to get a nice overview of the data. The table doesn't include nominal attributes. From the table it is evident that the attributes exhibit at large variety of value ranges. Based on this, we decided to normalize the attributes for use in the PCA.

	Mean	Variance	Standard Deviation
symboling	0.80	1.52	1.23
num-of-doors	3.16	0.97	0.99
wheel-base	98.92	37.66	6.14
length	174.33	154.91	12.45
width	65.89	4.55	2.13
height	53.87	5.71	2.39
curb-weight	2561.51	275975.55	525.33
num-of-cylinders	4.42	1.04	1.02
engine-size	128.12	1720.80	41.48
bore	3.33	0.07	0.27
stroke	3.25	0.10	0.31
compression-ratio	10.14	15.74	3.97
horsepower	103.48	1433.50	37.86
peak-rpm	5099.74	218536.20	467.48
city-mpg	25.33	40.59	6.37
highway-mpg	30.79	46.23	6.80
price	13285.03	65094229.49	8068.10

Table 1: Descriptive statistics of the attributes.

Visualization

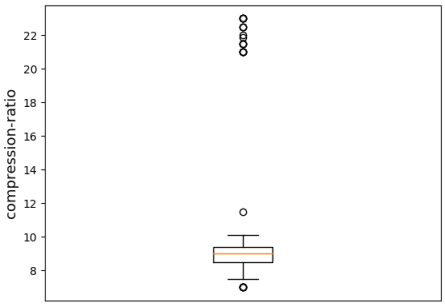
Using figures is a good way to learn new things about the data set. In this section continuous, non-nominal attributes are visualized and interpreted through, boxplots, histograms, correlation matrices, and scatter plots.
Due to the large amount of attributes, we have decided to show some of the plots that we found particularly interesting and leave the in the Appendix.

Boxplots

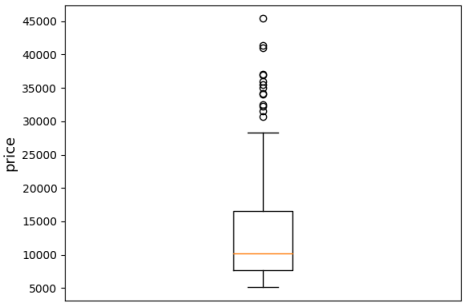
Boxplots make it easy to spot possible outliers in the data set. Looking at each boxplot, there are no data-points that lie out of the specified attribute intervals of the dataset. Even though several of the boxplots indicate outliers it does not imply faulty data or invalid values. The boxplot for the *Compression Ratio* attribute is a nice example of this (see figure 1a).
It's clear on this boxplot that there are several outliers - especially in the interval 20-23. This is not due to faulty data. Rather it is a side-effect of the data not being normally distributed. The values in *compression-ratio* seem to be clustered into two groups.
Summing up, the outliers will remain as valid values of the attributes and not be removed.

The *price* attribute is also very interestning (see figure 1b), because all the outliers lie above the median. This could indicate a heavy tail in the dataset and possibly that the data is lognormal distributed. Put differently, the data set contains a lot of mediocre cars, some super expensive luxury cars, but few super cheap cars.

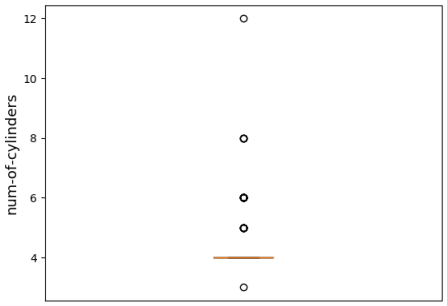
Outliers in the boxplot *Number of Cylinders* indicates whether the car has an extraordinary engine. It's also a special boxplot because most of the cars have four cylinders, so the 25th, 50th and the 75th percentile are placed on top of eachother (see figure 1c).
It's shown on the boxplot that at least one outlier has 12 cylinders. It's definitly extraordinary to have 12 cylinders in a car.



(a) Boxplot of Compression-Ratio



(b) Boxplot of Price



(c) Boxplot of Num-of-Cylinders

Figure 1

Histograms

Figure 2 shows the histograms for each of the non-nominal attributes. The data set contains roughly 200 observations, which is important to bear in mind when evaluating the data distribution. Some attributes like *length, width, height* and *peak-rpm* show signs of being normally distributed. Other attributes like *horsepower, engine-size, wheel-base* and *price* have a heavy right tail. Finally, there is *compression ratio* which, as mentioned before, has its observations clustered into two groups.

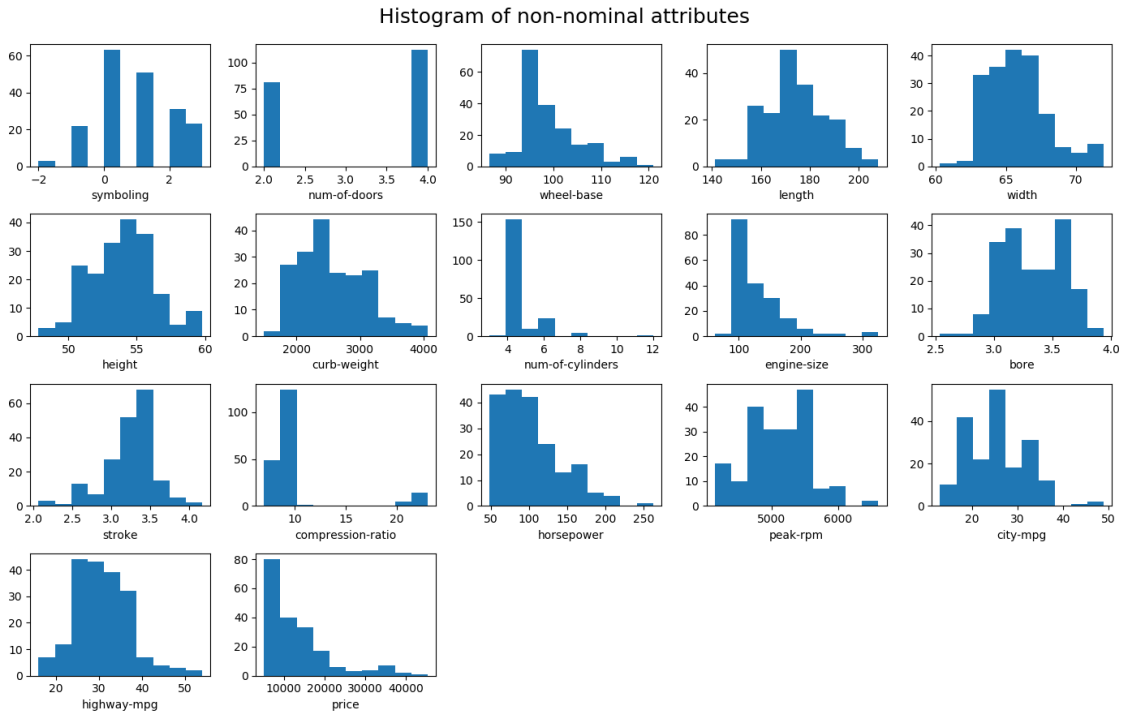


Figure 2: Histograms of attributes

Correlation

Scatter-plots are great for showing correlation between attributes. However, due to the amount of space required to plot 14x14 scatter-plots, we opted to move these to the Appendix. Instead, we will use a heatmap correlation matrix (figure 3) which roughly conveys the same information.

The correlation between *highway-mpg* and *city-mpg* is highly and positively correlated. This makes sense since the two attributes both measure gas efficiency in slightly different conditions.

Looking at the top left corner, which displays the correlations of *wheel-base, width, length, curb-weight* and *engine-size*, one can see that all of these are positively correlated. Intuitively this makes sense since they all express a measure of size and volume. Likewise, the same attributes are negatively correlated with the car mileage ultimately indicating that a larger and heavier car is less fuel efficient.
The correlation row/column for *price* is especially interesting because 10 out of 14 attributes are correlated with the price in some matter. The attribute is well correlated with *engine-size* and

horsepower, which is to be expected, since expensive cars typically have bigger engine, which in turn results in a lot of horsepower. It's also noticeable that expensive cars drive less miles per gallon than cheap cars do.
Stroke and *compression-ratio* is very badly correlated with any of the other attributes. The *compression ratio* is the ratio between the volume of the cylinder and the combustion chamber, when the piston is at the bottom of it's stroke, and the volume of the combustion chamber when the piston is at the top of it's stroke. *Stroke* is the distance traveled by the piston each cycle. Intuitively *stroke* and *combustion-ratio* should be more correlated than 0.2.

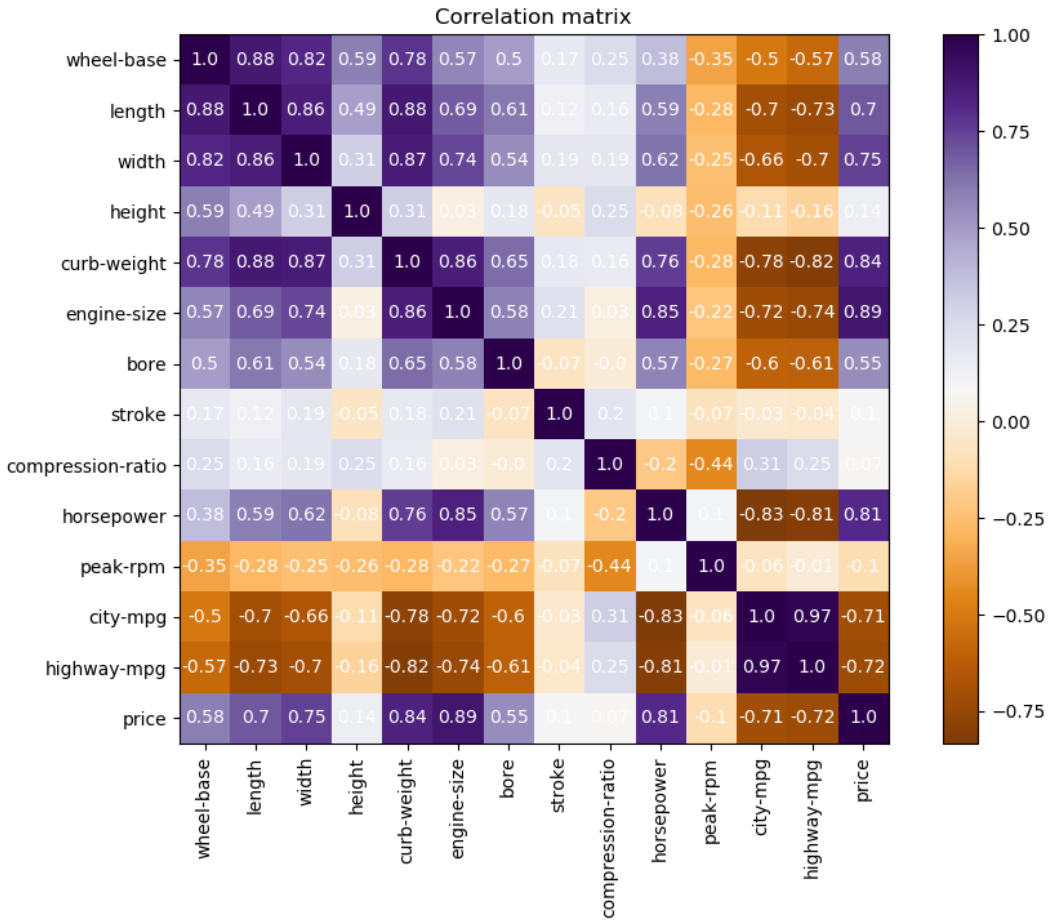


Figure 3: Correlation of attributes

PCA

After 1-out-Of-K-encoding and normalization of the data, we did principle component analysis through eigenvalue decomposition. With 1-out-Of-K-encoding, we ended up with more than 60 columns of data. Using roughly the first 15 principle components, we were able to explain 90% of the variance in the data (see figure 4).

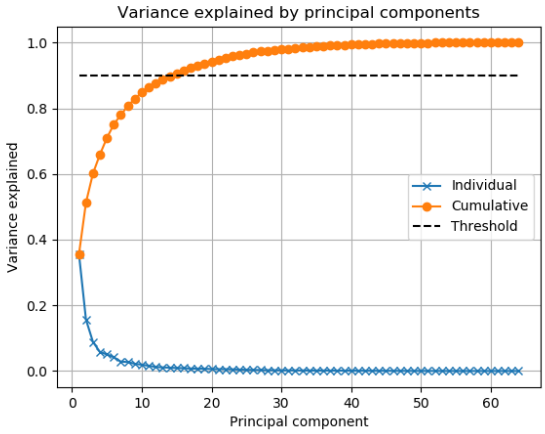


Figure 4: Individual and cumulative explained variance of the principle components.

Due to the large amount of dimensions, we decided to show the 16 attributes with the highest amplitude for the first three principle components (see figure 5). Looking at the largest values of the principle components it's hard to give a definite answer for what each of them may describe. The first principle component seems to distinguish between small and large cars. A small car will usually have a small engine, width and wheelbase which is what the plot shows. For the first two principle components, *bore* seems to be the most important attribute.

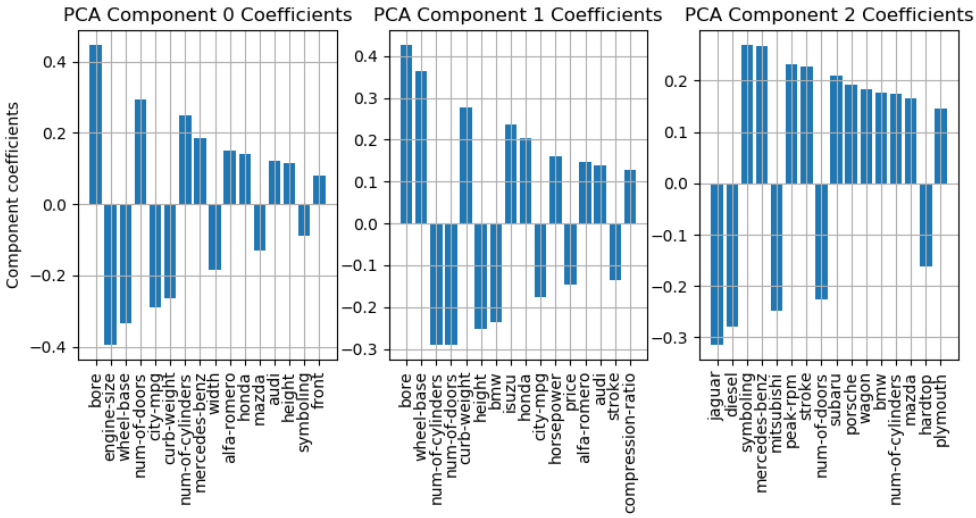


Figure 5: The 16 most important attributes of the first three principle components

In figure 6 is shown the projection of the data onto the first three principle components. Without classes and a clear definition of what each principle component describes, it's hard to draw any conclusions from the plots.

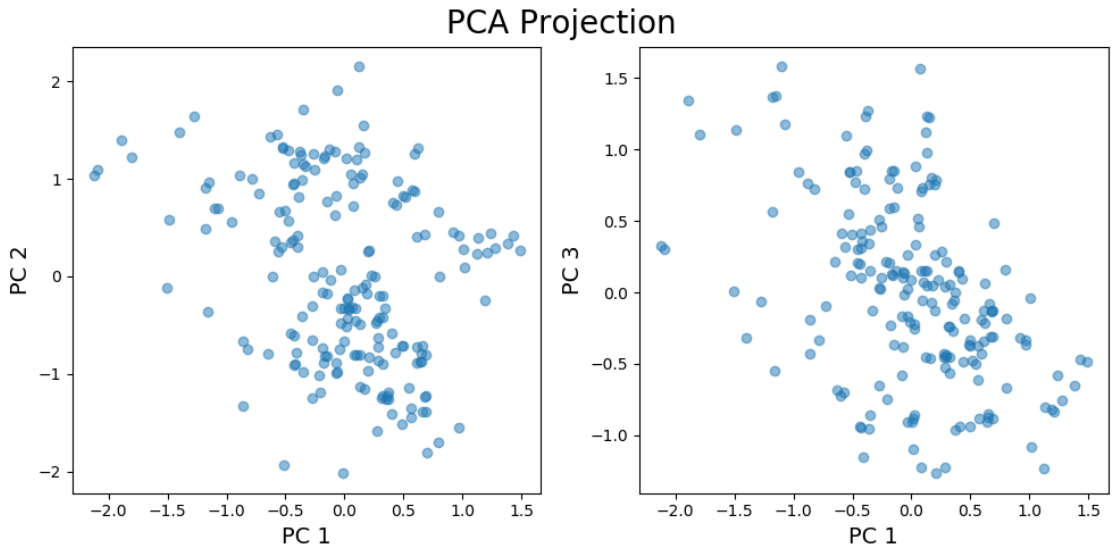


Figure 6: Projection of the data onto the first three principle components

Discussion

Our data contained some empty values that belonged mostly to the attribute *normalized-losses*. Except from that we found no faulty recordings in the data. Our PCA showed that roughly 15 principle components will give an explained variance above 90%. Machine-learning wise, we are a bit worried that the data set is too high-dimensional (64) compared to the number of observations (193). Hopefully, using the principle components we will be able to reduce the dimensionality. Our correlation matrix of the continuous attributes also showed that some attributes were highly correlated such as *city-mpg* and *highway-mpg*, which is a sign in favor of using PCA. A reason to why compression ratio and stroke are not correlated more than 0.2, could be the that compression ratio and stroke have different interpretations for different types of engines.

Appendix

Workload

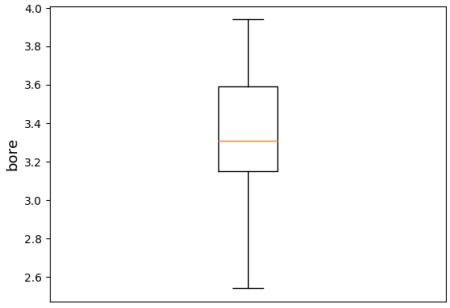
Section	Erik Gylling	Christian Hinge
Introduction	$\frac{1}{2}$	$\frac{1}{2}$
Overview of Data	$\frac{1}{2}$	$\frac{1}{2}$
Visualization	$\frac{1}{2}$	$\frac{1}{2}$
Discussion	$\frac{1}{2}$	$\frac{1}{2}$

Attributes

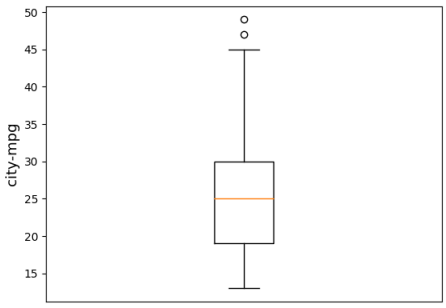
- **symboling** is a discrete ordinal attribute with the values: -3, -2, -1, 0, 1, 2, 3.
  The degree to which the auto is more risky than its price indicates. -3 indicates that the car is pretty safe while a value of 3 indicates a highly risky car.
- **normalized-losses** is a continuous ratio attribute with values from: 65 to 256.
  Represents the average loss per car per year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...).
- **num-of-doors** is a discrete ratio attribute with the values: four, two.
  This attribute specifies the number of doors in the car.
- **wheel-base** is a continuous ratio attribute with values from: 86.6 120.9.
  This attribute specifies the distance between front wheels and rear wheels of a car.
- **length** is a continuous ratio attribute with values from: 141.1 to 208.1.
  This attribute specifies the length of the car.
- **width** is a continuous ratio attribute with values from: 60.3 to 72.3.
  This attribute specifies the width of the car.
- **height** is a continuous ratio attribute with values from: 47.8 to 59.8.
  This attribute specifies the height of the car.
- **bore** is a continuous ratio attribute with values from: 2.54 to 3.94.
  This attribute specifies the diameter of the cylinders.
- **stroke** is a continuous ratio attribute with values from: 2.07 to 4.17.
  This attribute specifies the distance travelled by the piston in each cycle.
- **compression-ratio**is a continuous ratio attribute with values from: 7 to 23.
  This attribute specifies ratio between the volume of the cylinder and combustion chamber, when the piston is at the bottom of its stroke, and the volume of the combustion chamber when the piston is at the of its stroke..
- **horsepower** is a continuous ratio attribute with values from: 48 to 288.
  This attribute specifies the amount of horsepower of the car.
- **peak-rpm** is a continuous ratio attribute with values from: 4150 to 6600.
  This attribute specifies the maximum revolutions per minute the engine of the car can perform.
- **highway-mpg** is a continuous ratio attribute with values from: 16 to 54.
  This attribute specifies how many miles the car drives per gallon of fuel on the highway.

- **make** is a discrete nominal attribute, with the categories: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo.
  This attribute specifies the make of the car.
- **fuel-type** is a discrete nominal attribute, with the categories: diesel, gas.
  This attribute specifies the type of fuel.
- **aspiration** is a discrete nominal attribute, with the categories: std, turbo.
  Standard engines rely solely on atmospheric pressure for air-intake while engines with a turbo can force air into the engine.
- **drive-wheels** is a discrete nominal attribute, with the categories: 4wd, fwd, rwd.
  This attribute specifies form which wheels the car gets its propulsion.
- **engine-location** is a discrete nominal attribute, with the categories: front, rear.
  This attribute specifies where the engine is located.
- **engine-type** is a discrete nominal attribute, with the categories: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
  This attribute specifies the engine type.
- **fuel-system** is a discrete nominal attribute, with the categories: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
  This attribute specifies which fuel system the car use.

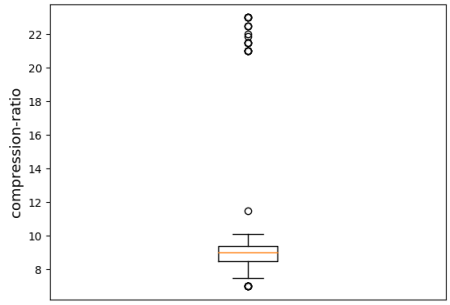
Plots



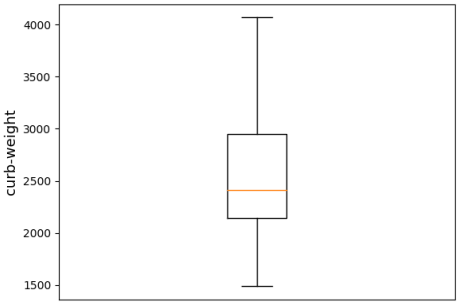
(a) Boxplot for Bore



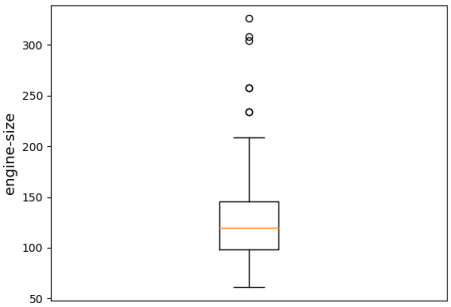
(b) Boxplot for City-mpg



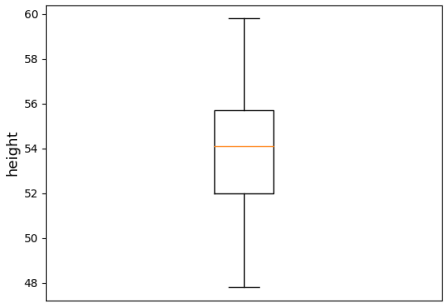
(c) Boxplot for Compression-Ratio



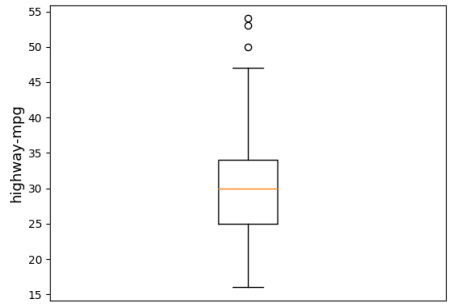
(d) Boxplot for Curb-weight



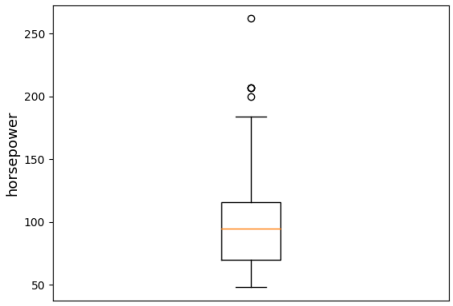
(e) Boxplot for Engine-size



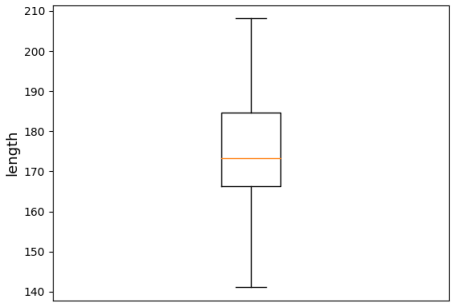
(f) Boxplot for Height



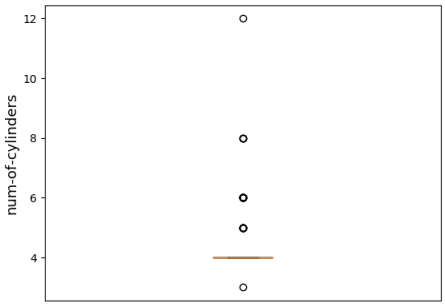
(g) Boxplot for Highway-mpg



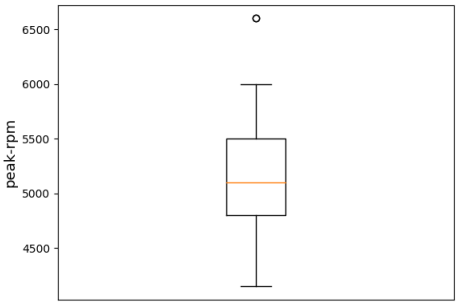
(h) Boxplot for Horsepower



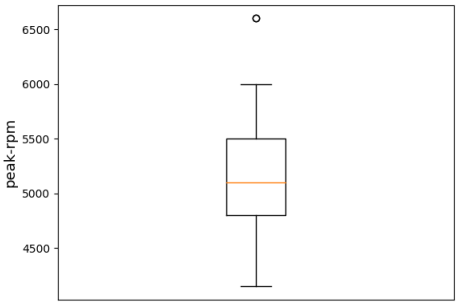
(i) Boxplot for Length



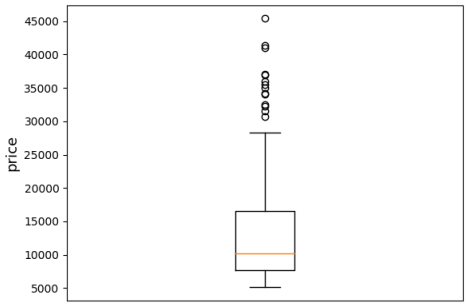
(j) Boxplot for Num-of-Cylinders



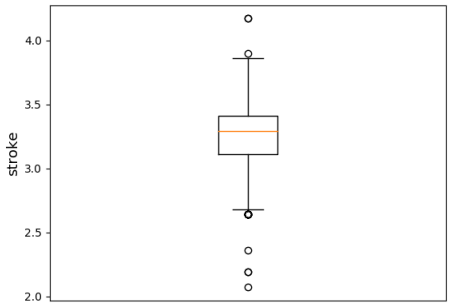
(k) Boxplot for Num-of-Doors



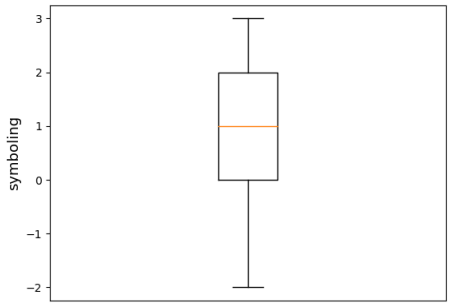
(l) Boxplot for Peak-rpm



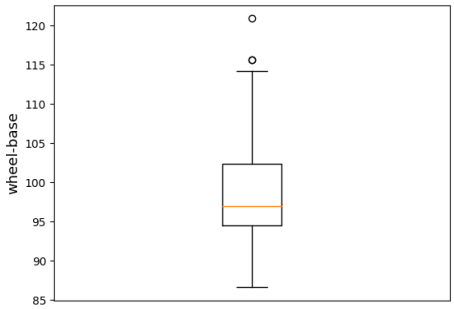
(m) Boxplot for Price



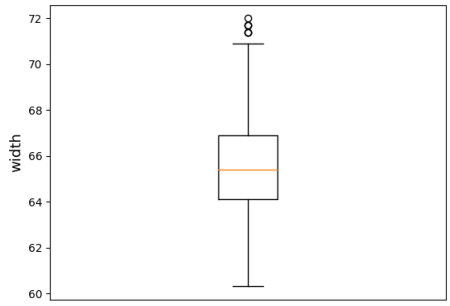
(n) Boxplot for Stroke



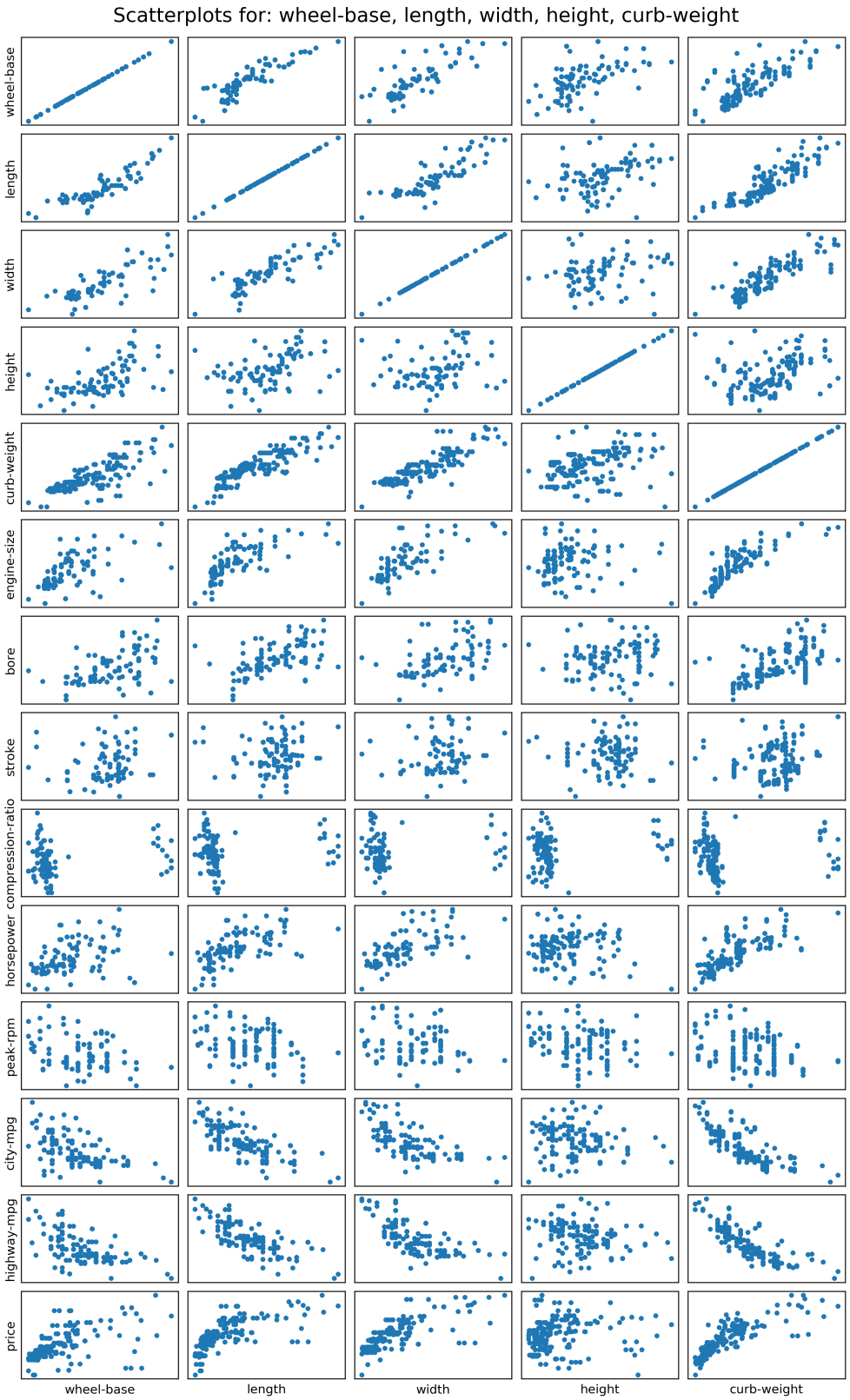
(o) Boxplot for Symboling



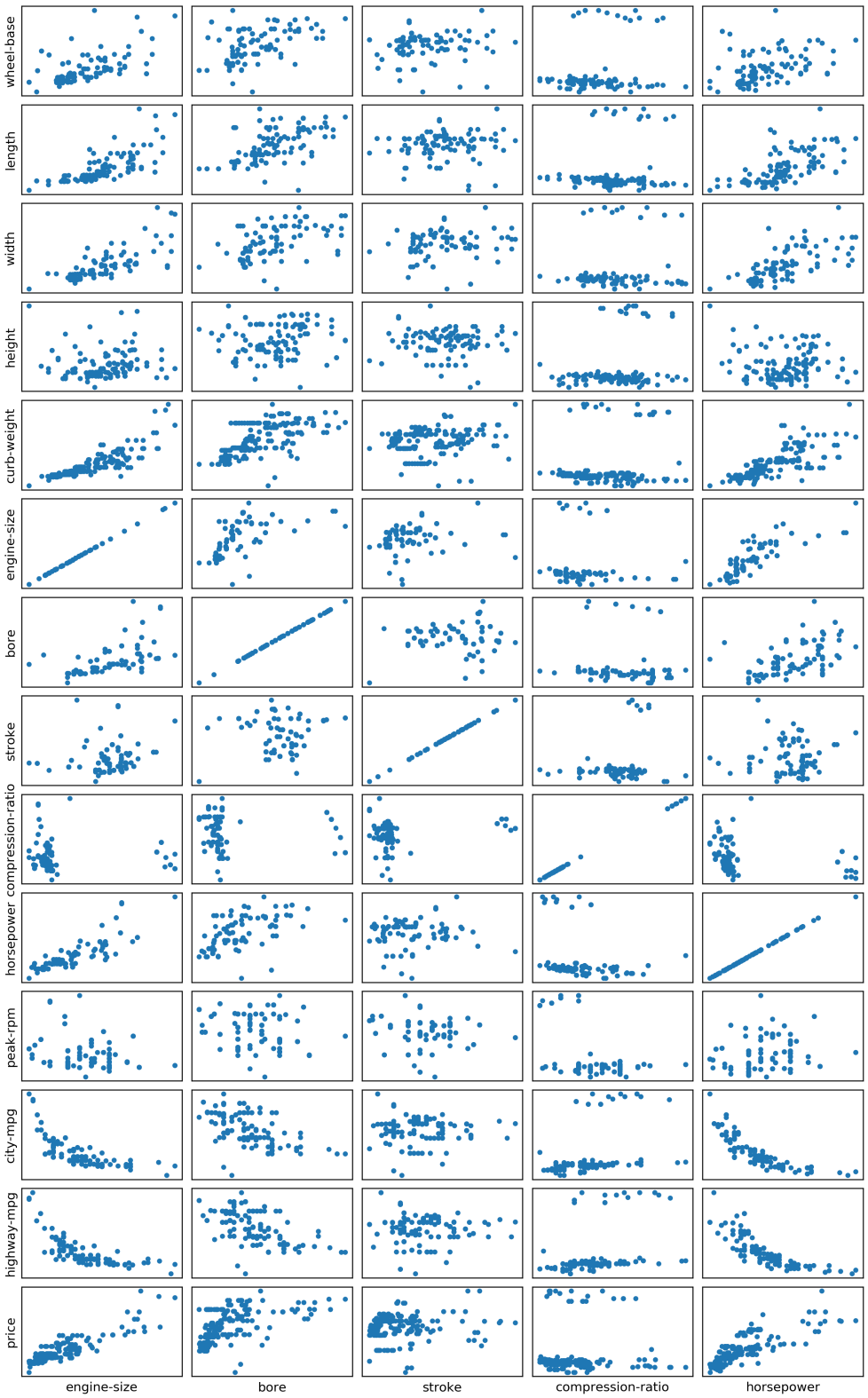
(p) Boxplot for Wheel-base

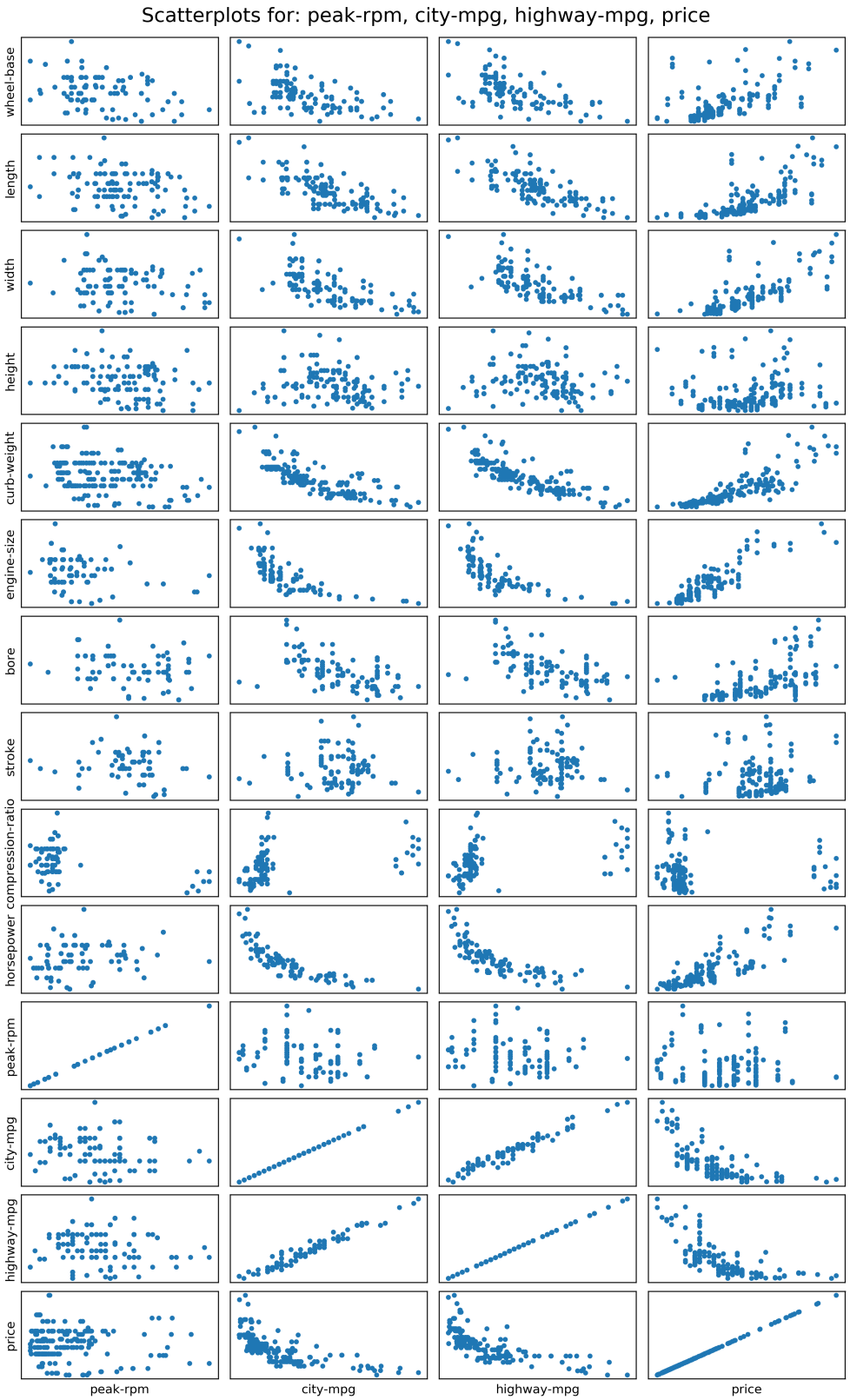


(q) Boxplot for Width


Scatterplots for: wheel-base, length, width, height, curb-weight
wheel-base
length
width
height
curb-weight
engine-size
bore
stroke
compression-ratio
horsepower
peak-rpm
city-mpg
highway-mpg
price
wheel-base
length
width
height
curb-weight

Scatterplots for: engine-size, bore, stroke, compression-ratio, horsepower




Scatterplots for: peak-rpm, city-mpg, highway-mpg, price
wheel-base
length
width
height
curb-weight
engine-size
bore
stroke
compression-ratio
horsepower
peak-rpm
city-mpg
highway-mpg
price
peak-rpm
city-mpg
highway-mpg
price