ECE 580 – Spring 2020 – Homework 2
Andre Costa

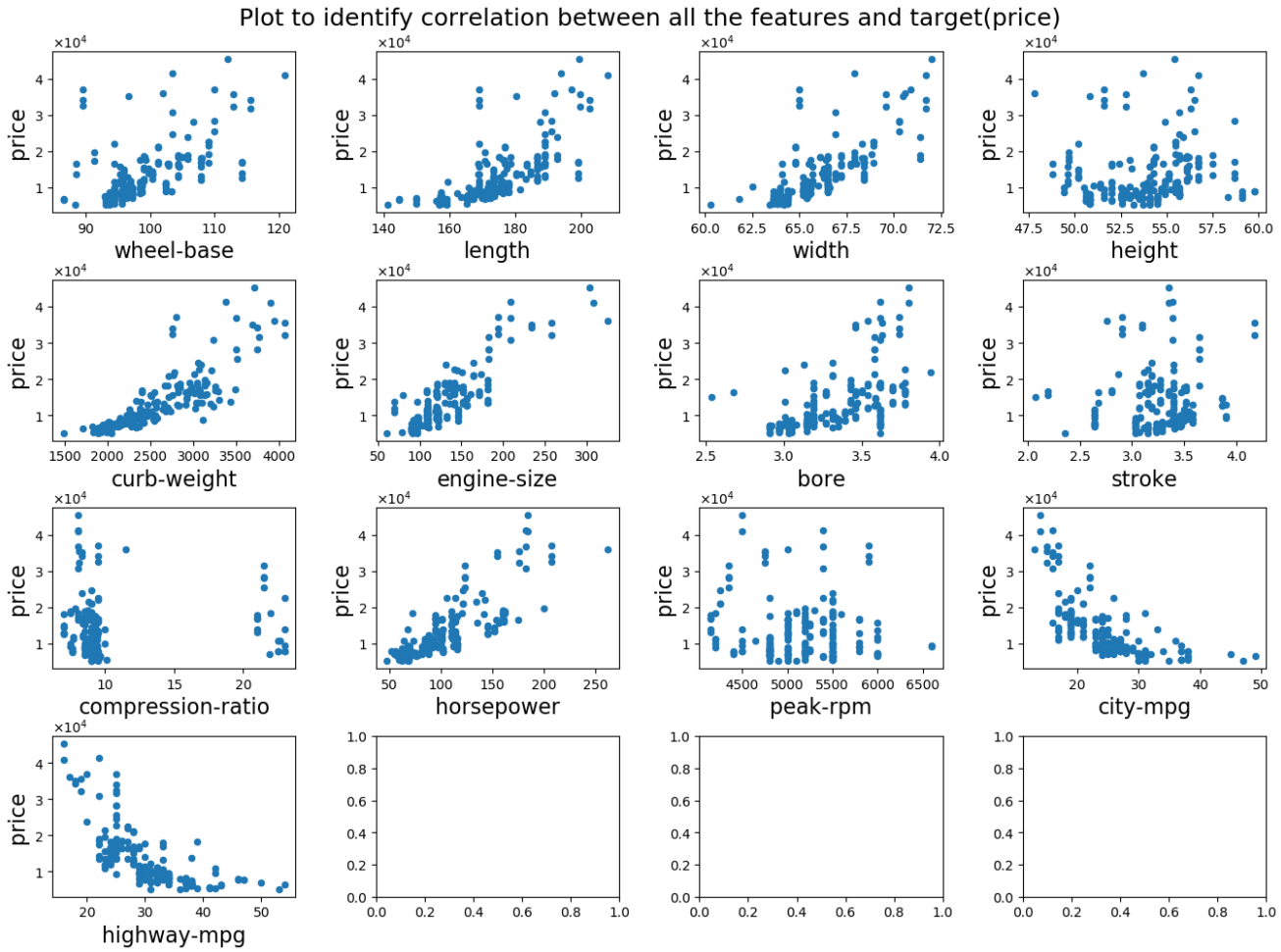
Part 1: Data Preparation and Exploration

1 - a) I used Python's library Pandas to clean the data set. It has some very useful functions to treat tabular data. The first step was to read the csv data from the text file. This is easily done by Pandas' read_csv function. In this step, I also set the NaN(invalid or missing data) value to '?'. This would make data cleaning very simple. In this step, I obtained a total of 205 data points, with 25 features each.

The second step was to clean the data. First, I removed all the non-continuous features by deleting column by column (this was not very efficient, I will try to find a better way to do this in next homeworks). In this step, 12 features (columns) were removed from the data set. Next, I removed all data points that contained missing values. This was very easy, because I had already defined '?' as a NaN value. So, I only had to use the dropna function from pandas to get rid of data points with '?' in the price column. This step removed 4 data points.

After that, I reorganized the column's names and row's ordering and finally my data set, consisting of 201 data points and 13 features, was ready to be analyzed.

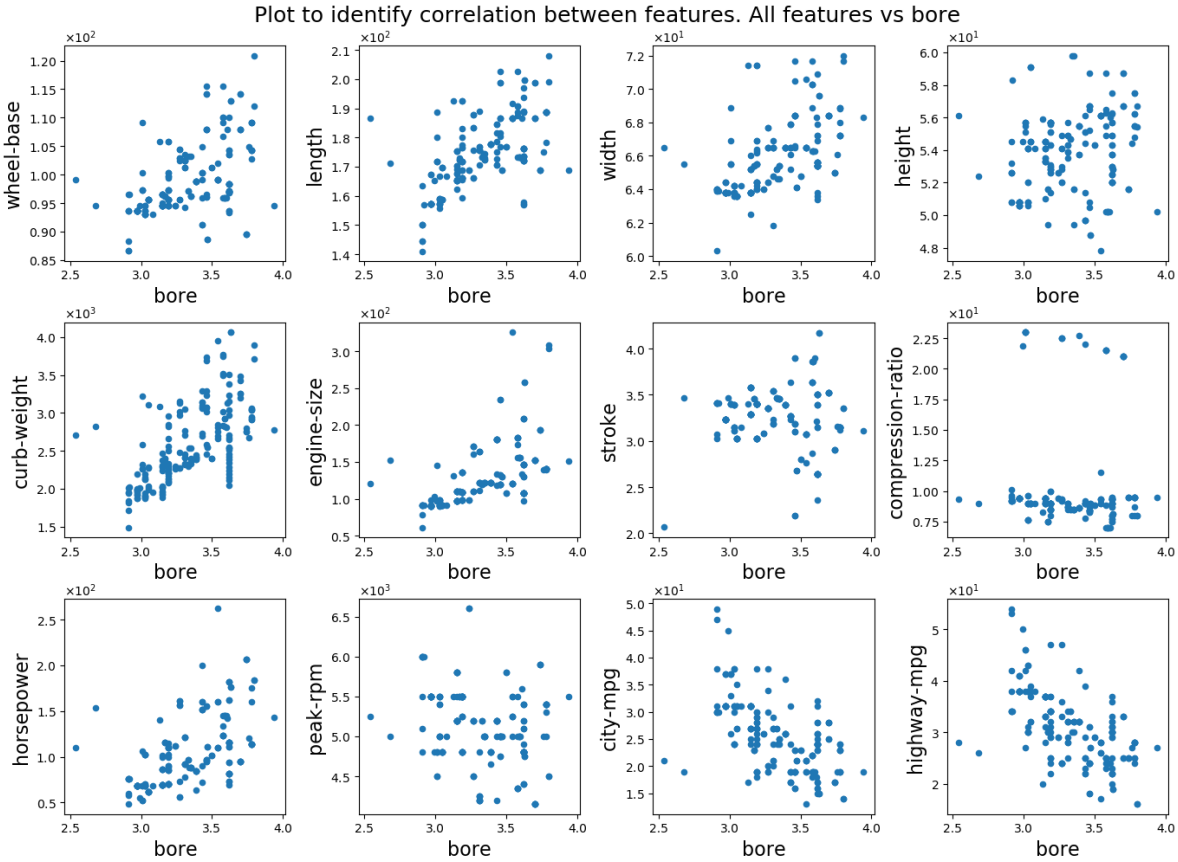
b) This plot was generated with pandas and matplotlib libraries. It shows the relation between each of the features and the target variable (car's price).



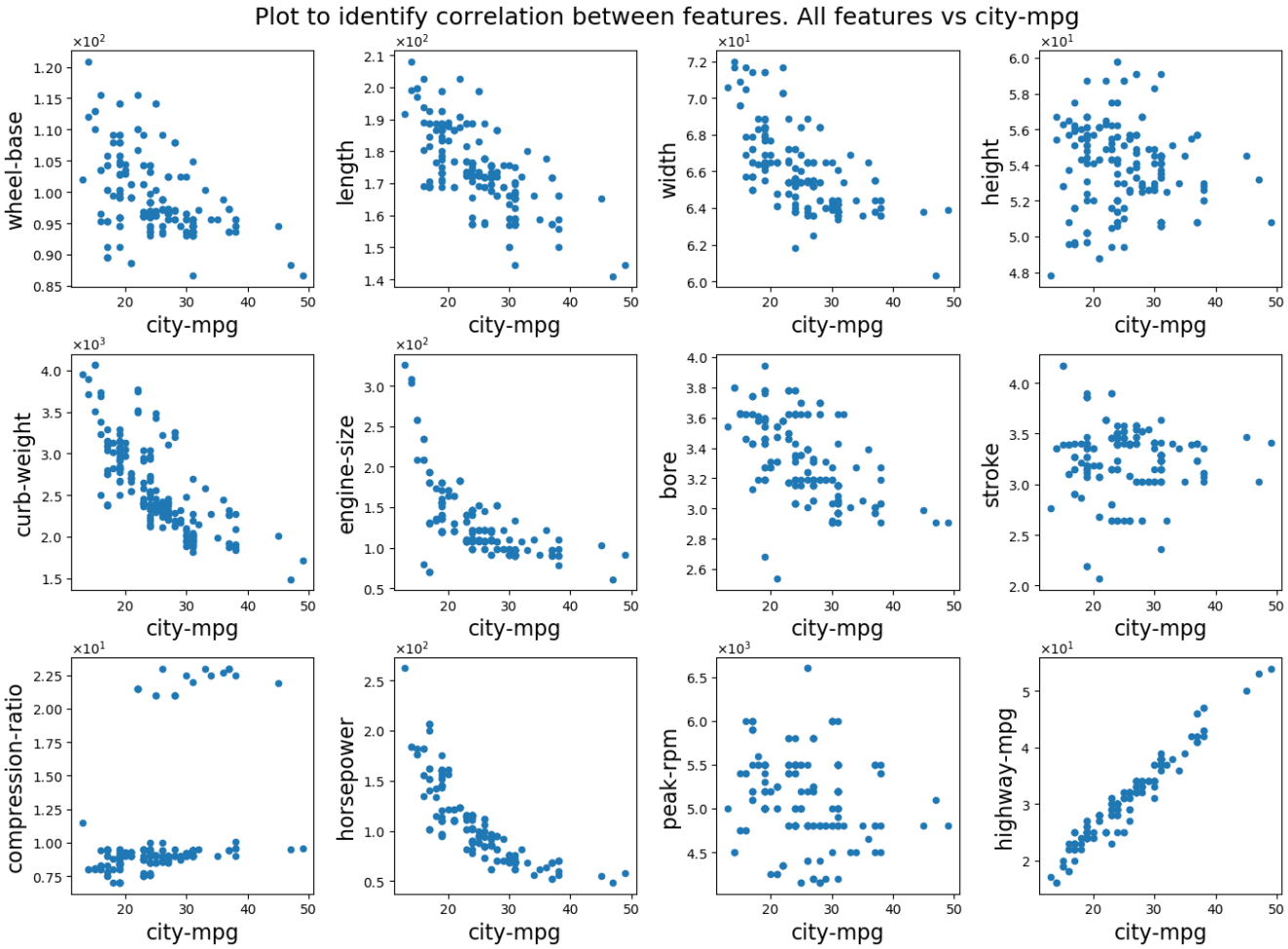
c) We can clearly see some trends in the data, but also some features that seen totally unrelated to the target. Wheel-base, height, stroke, compression-ratio and peak-rpm all seem to have no influence in price. On the other hand, length, width, curb-weight, engine-size, bore and horse power all seem to be positively correlated to the price, while city-mpg and highway-mpg seem to be negatively correlated. By simply looking at the plots, I can also see some type of quadratic relation between length, width, curb-weight and bore to the price. Whereas for engine-size and horse-power, this relation seem more linear. Therefore, we can separate the features in 4 groups:

I) Uncorrelated to price: wheel-base, height, stroke, compression-ratio and peak-rpm.
II) Linear relation: engine-size and horse-power
III) Quadratic relation: length, width, curb-weight and bore
IV) Inverse relation: city-mpg and highway-mpg

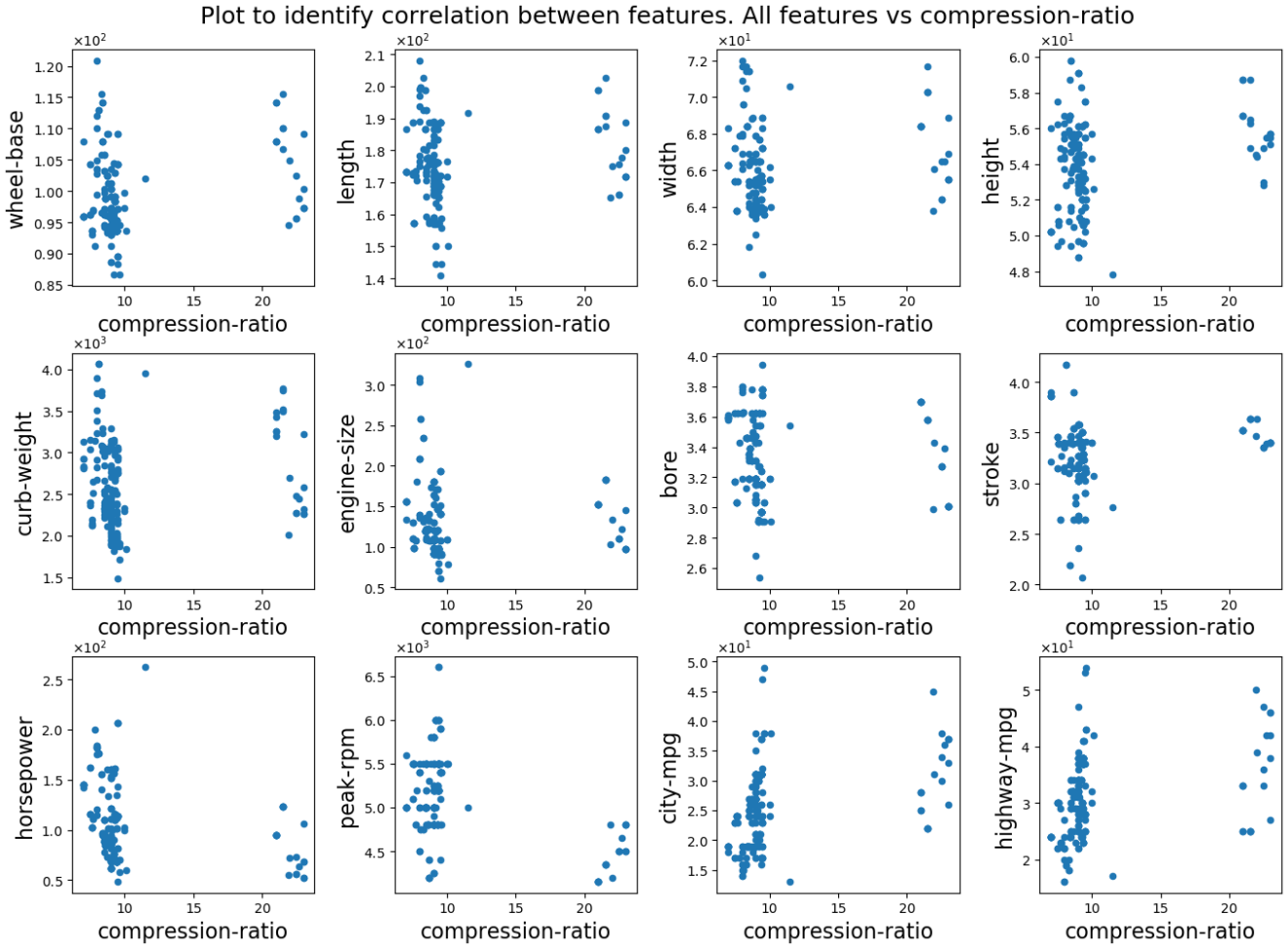
d)



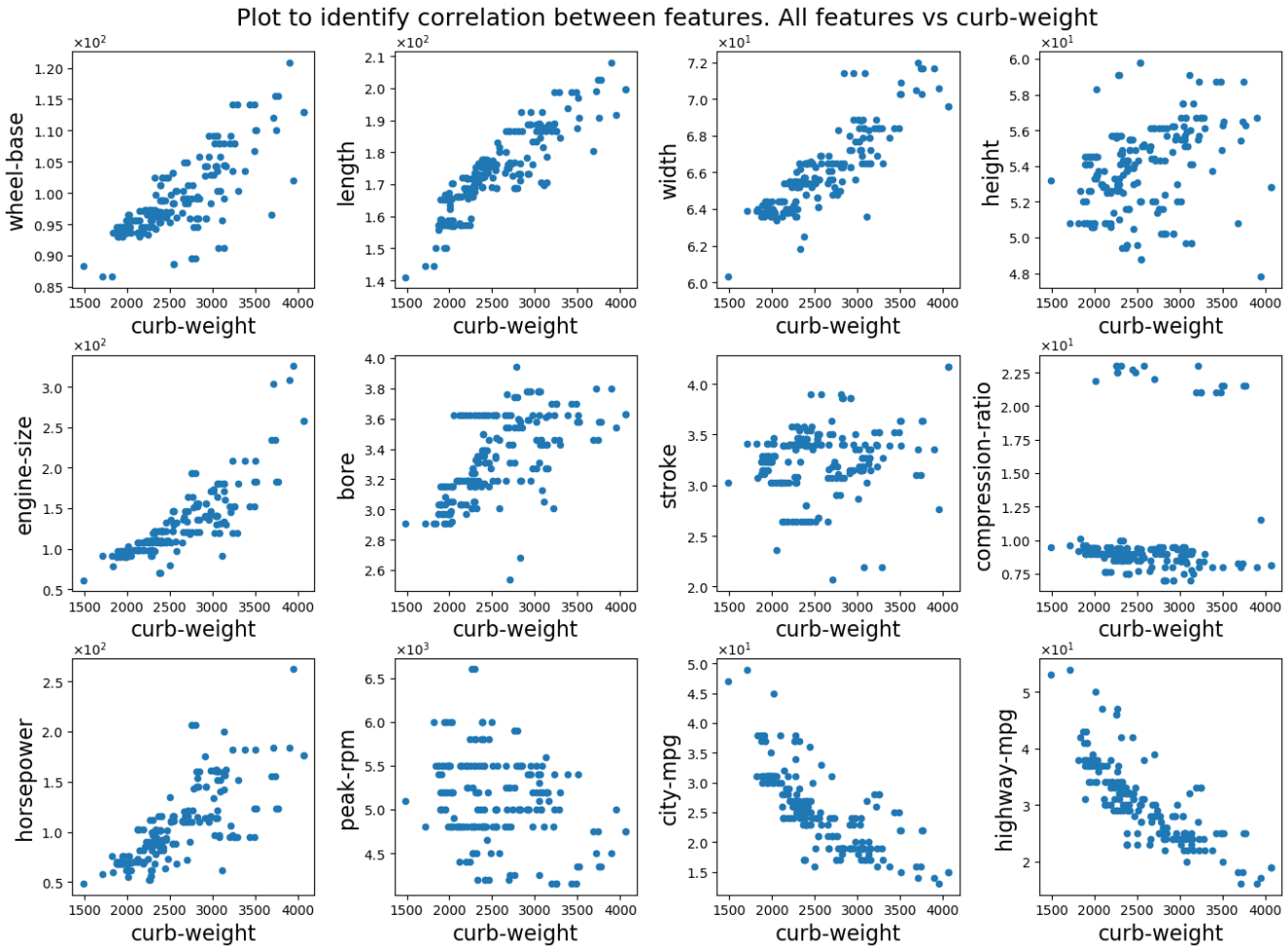
Looking at these plots, we don't see a very strong correlation of bore with any of the other features, but we can see that there is some correlation, specially with curb-weight, length, engine size and horsepower. We can also see a negative correlation with city-mpg and highway-mpg.



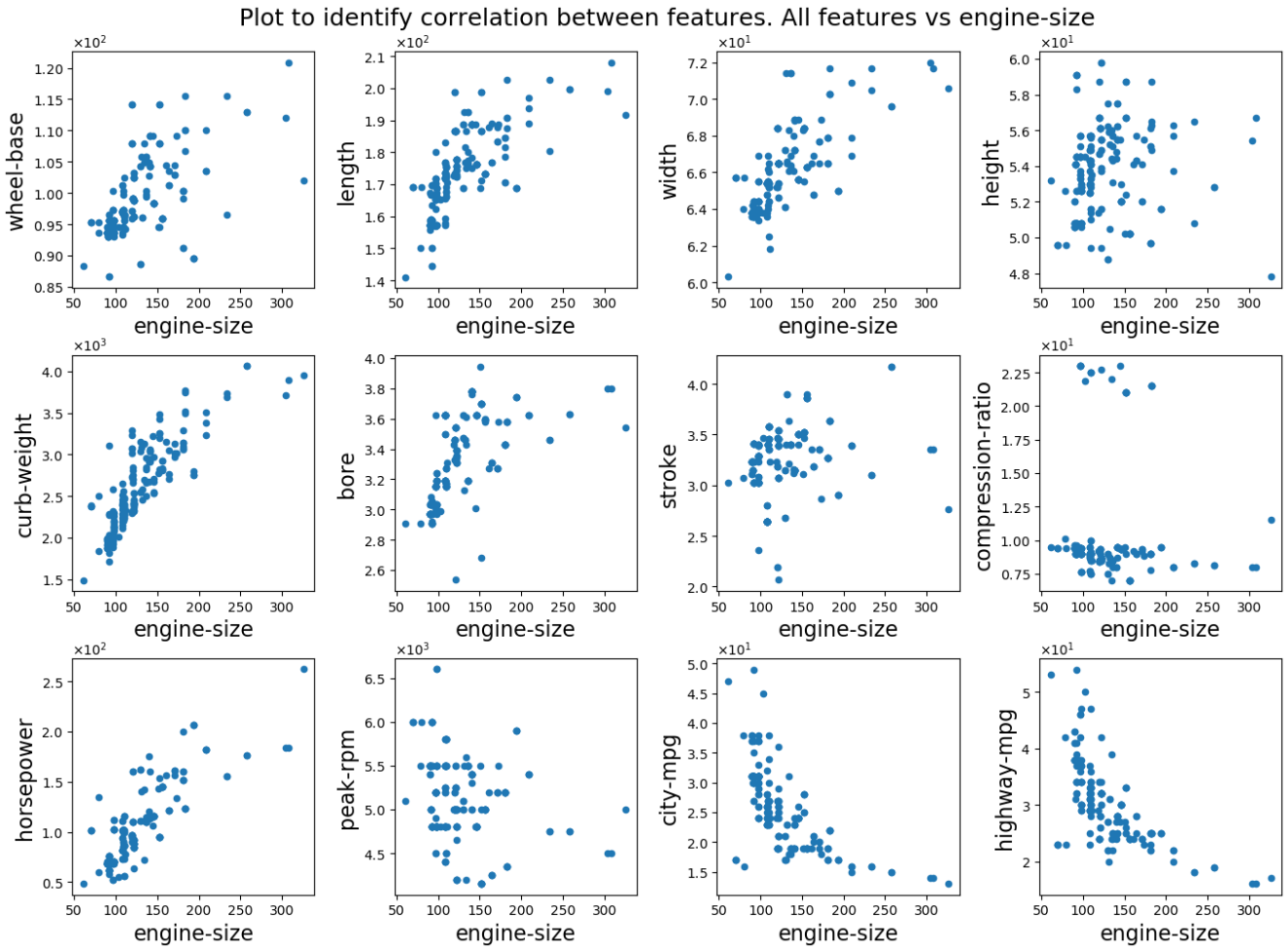
For city-mpg, we can see a very strong correlation with highway-mpg, which is expected and a strong negative correlation with horse-power and engine size. There is also some negative correlation with curb-weight but it doesn't seem to be very strong.



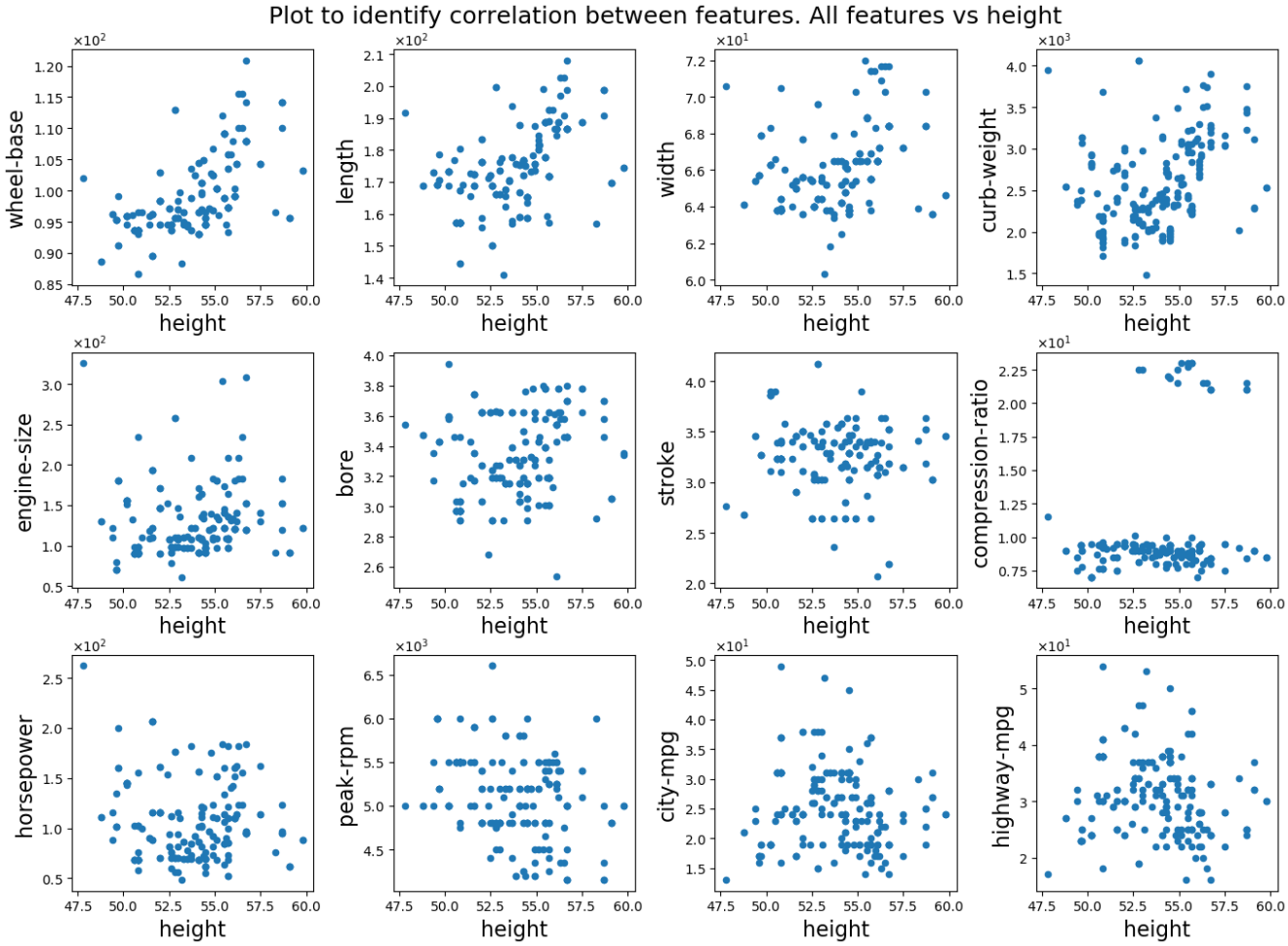
Compression-ratio doesn't seem to be correlated with any of the other features. It also didn't show any relation to the target value (recall from item c).



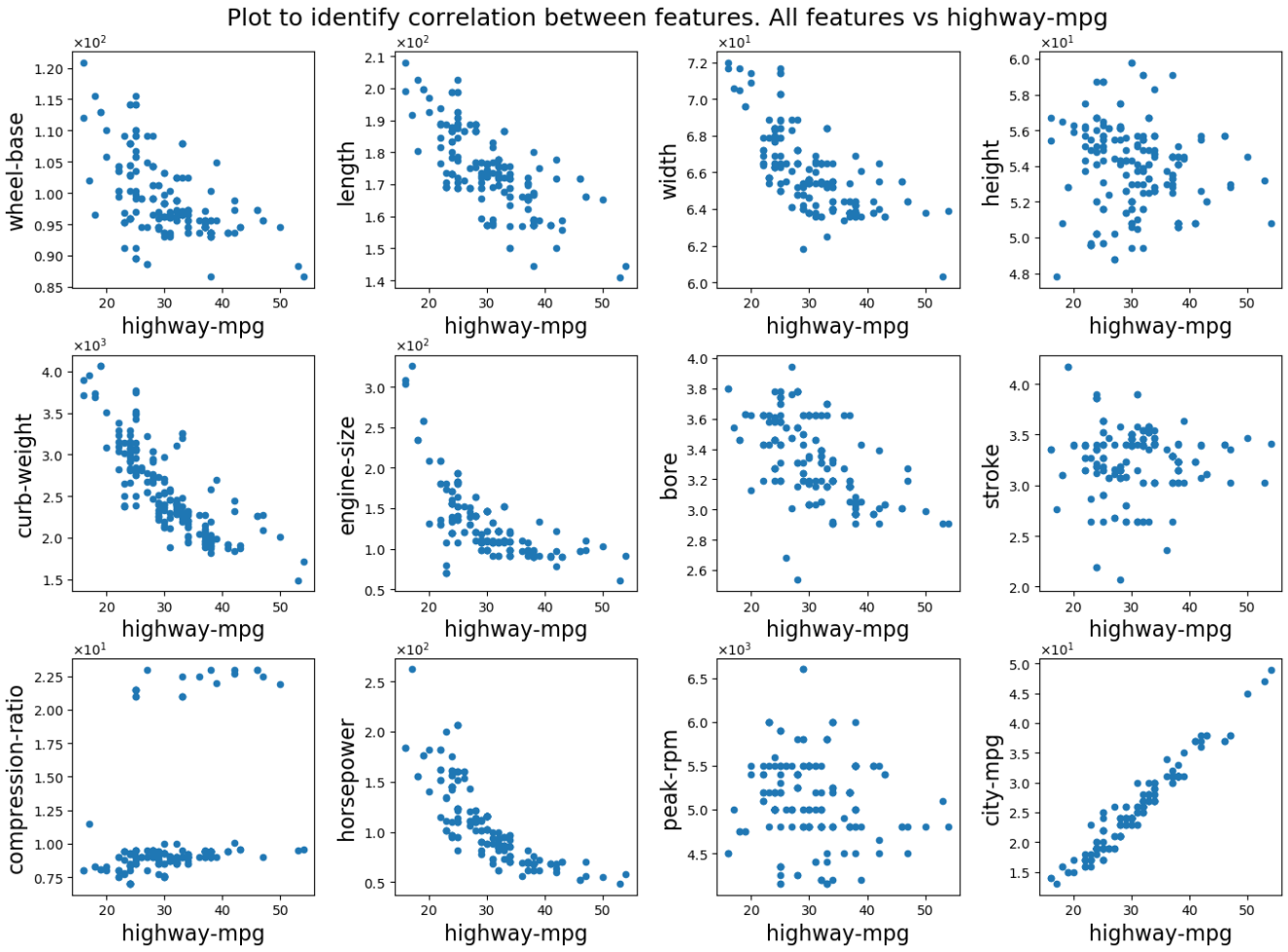
Curb-weight shows considerable correlation with length and width, and also, some correlation with engine-size and horsepower. It also shows negative correlation with city and highway-mpg.



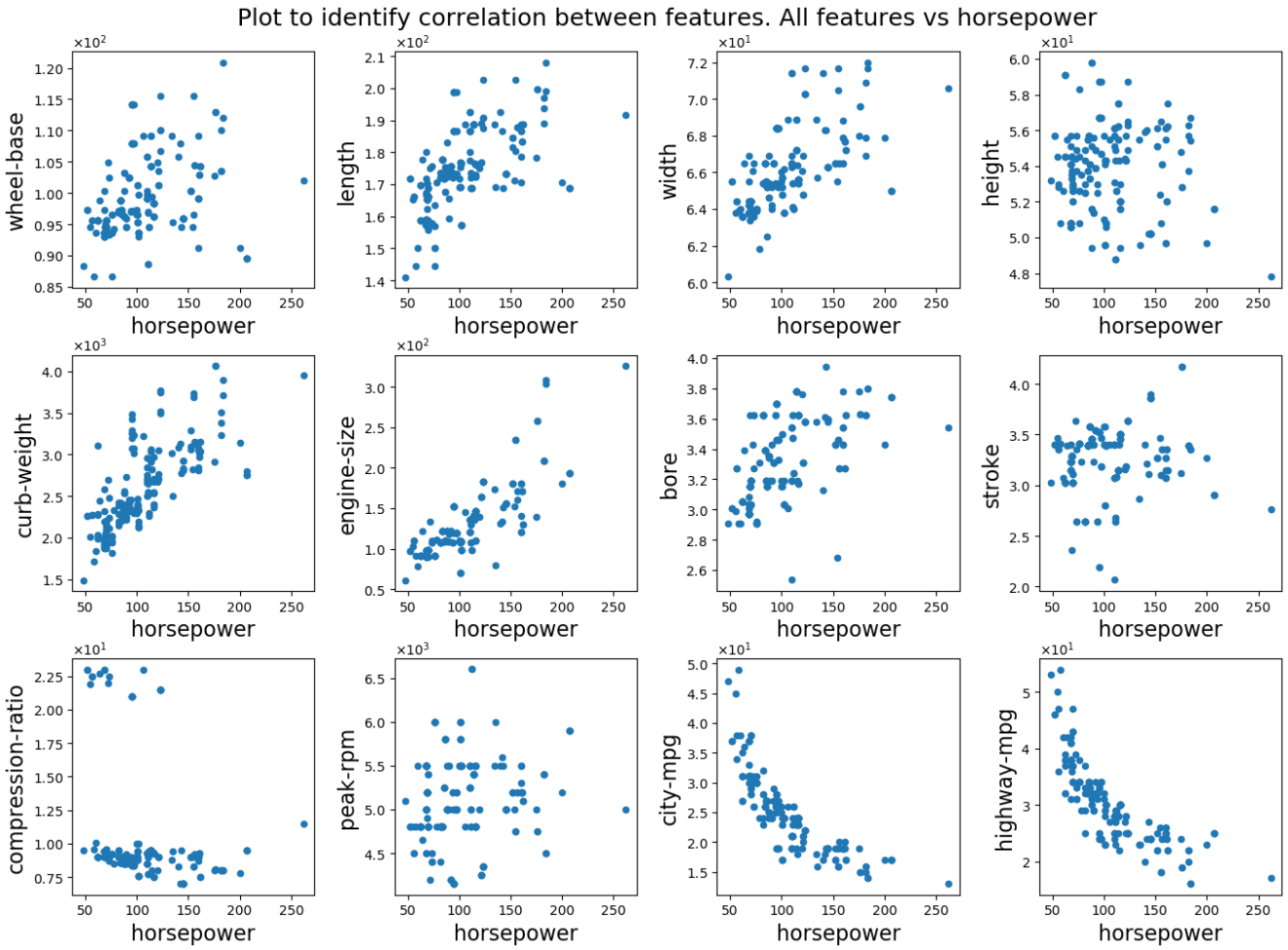
Engine-size has some considerable correlation with curb-weight and horsepower, and some strong negative correlation with city-mpg and highway-mpg. There is also some weak correlation with length and width.



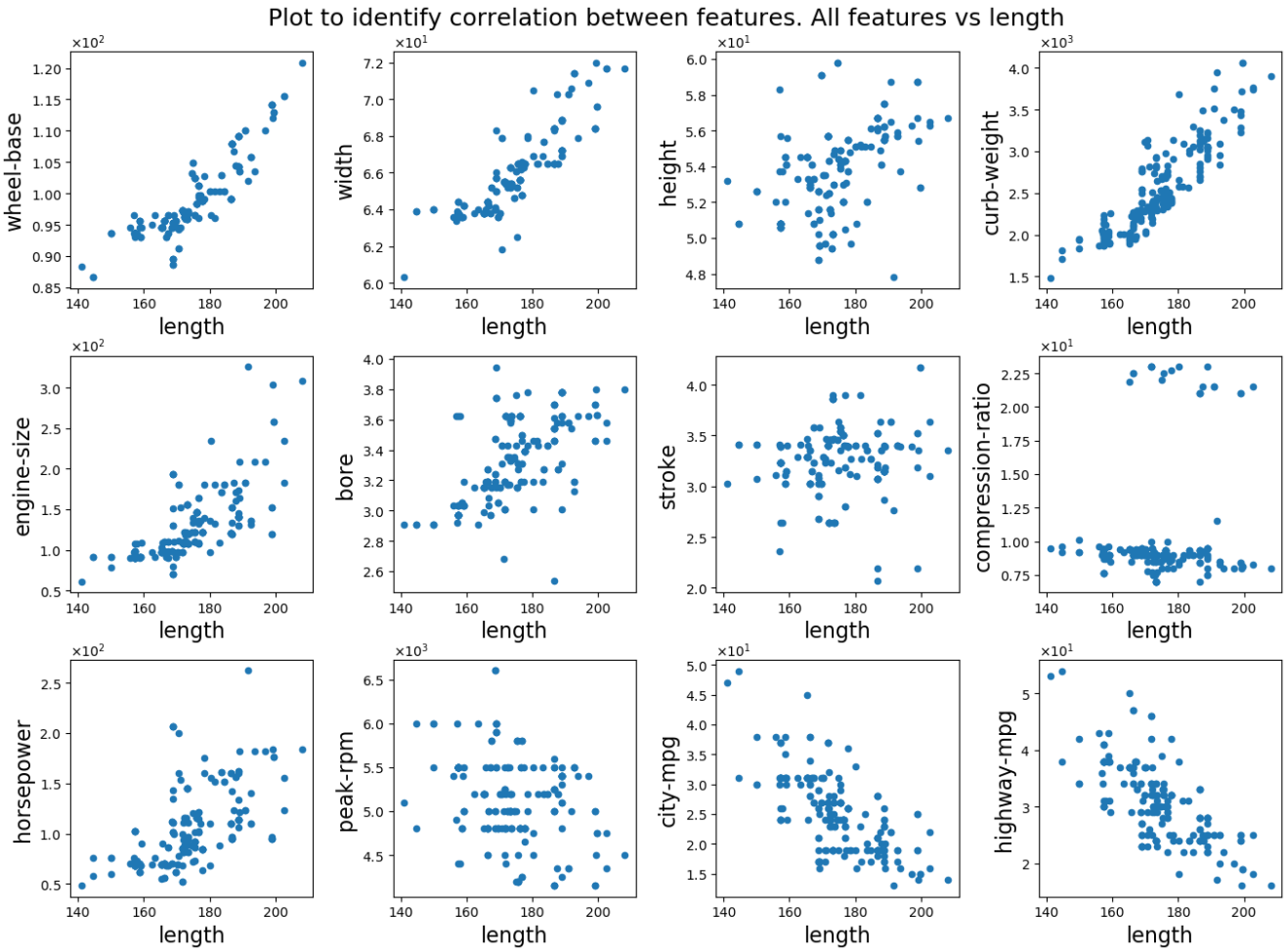
Height doesn't seem to be correlated with any of the other features. It also didn't show any relation to the target value (recall from item c).



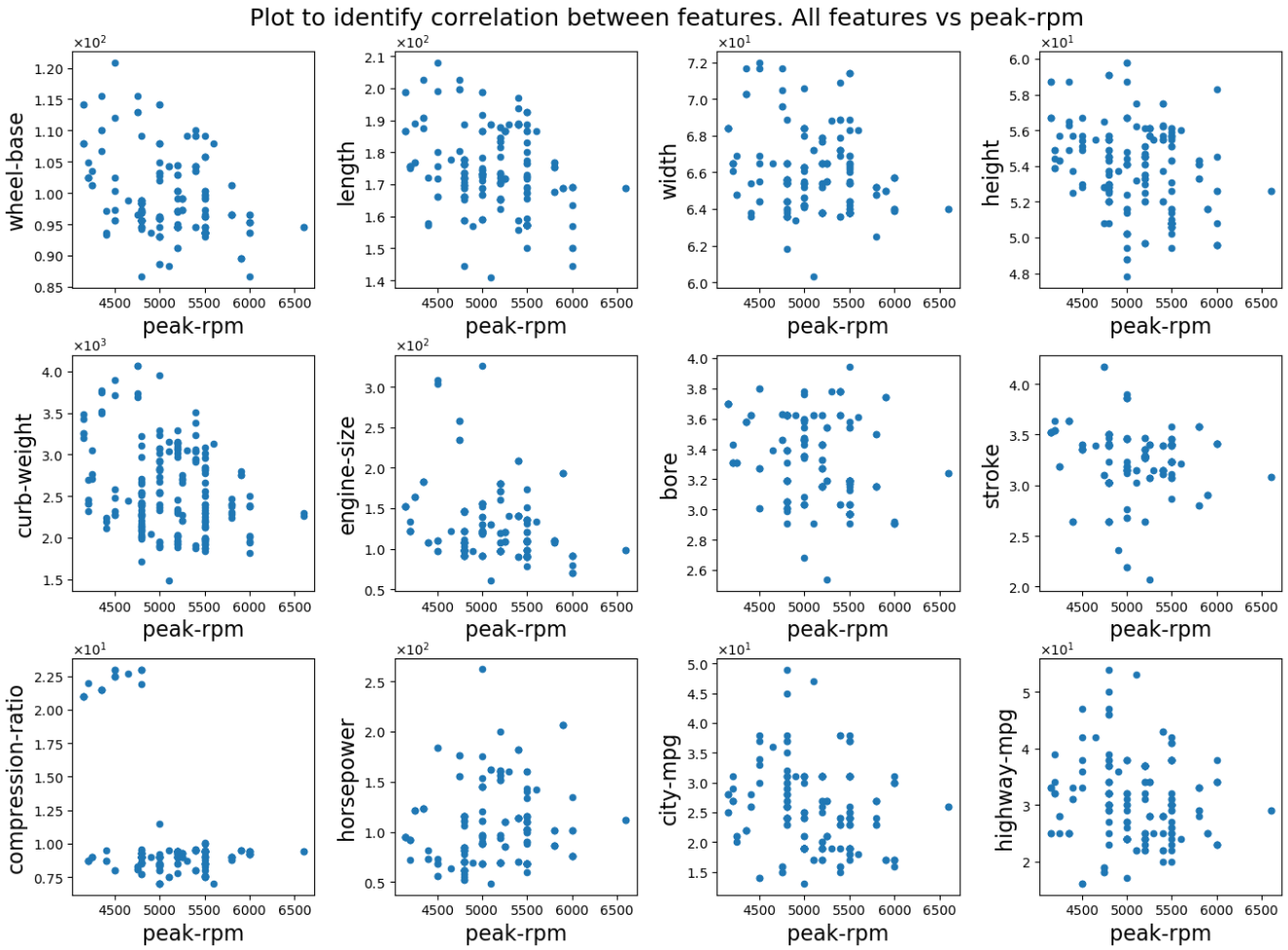
For highway-mpg, we can see a very strong correlation with city-mpg, which is expected, and a strong negative correlation with horse-power, engine size and curb-wight. There is also some negative correlation with length and width but it doesn't seem to be very strong.



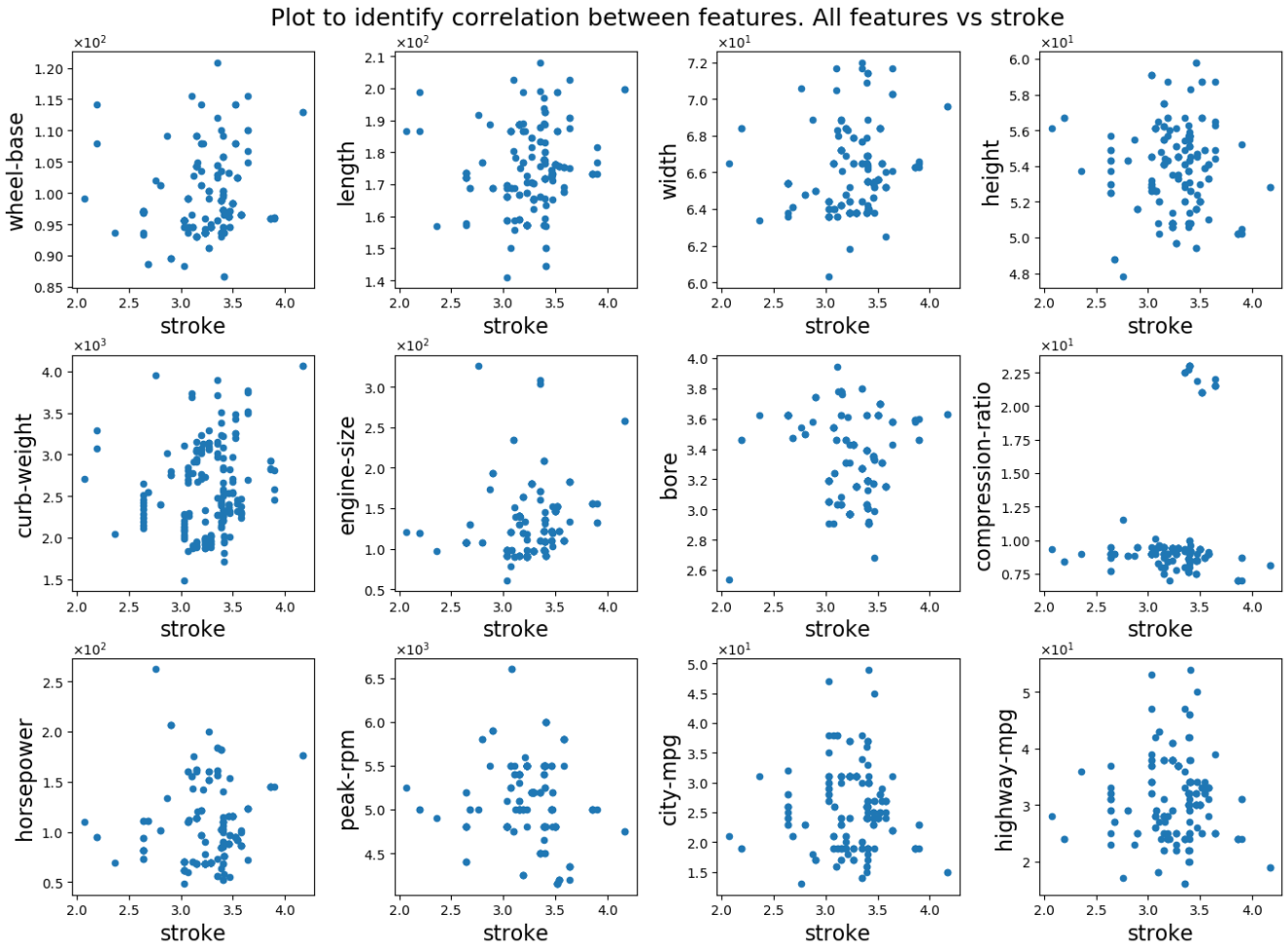
Here, we can see the strong negative correlation between horsepower and city and highway-mpg. We can also see a considerable correlation between horsepower and curb-weight and engine size. Finally, there is some correlation between horsepower and length and width but it seems very weak.



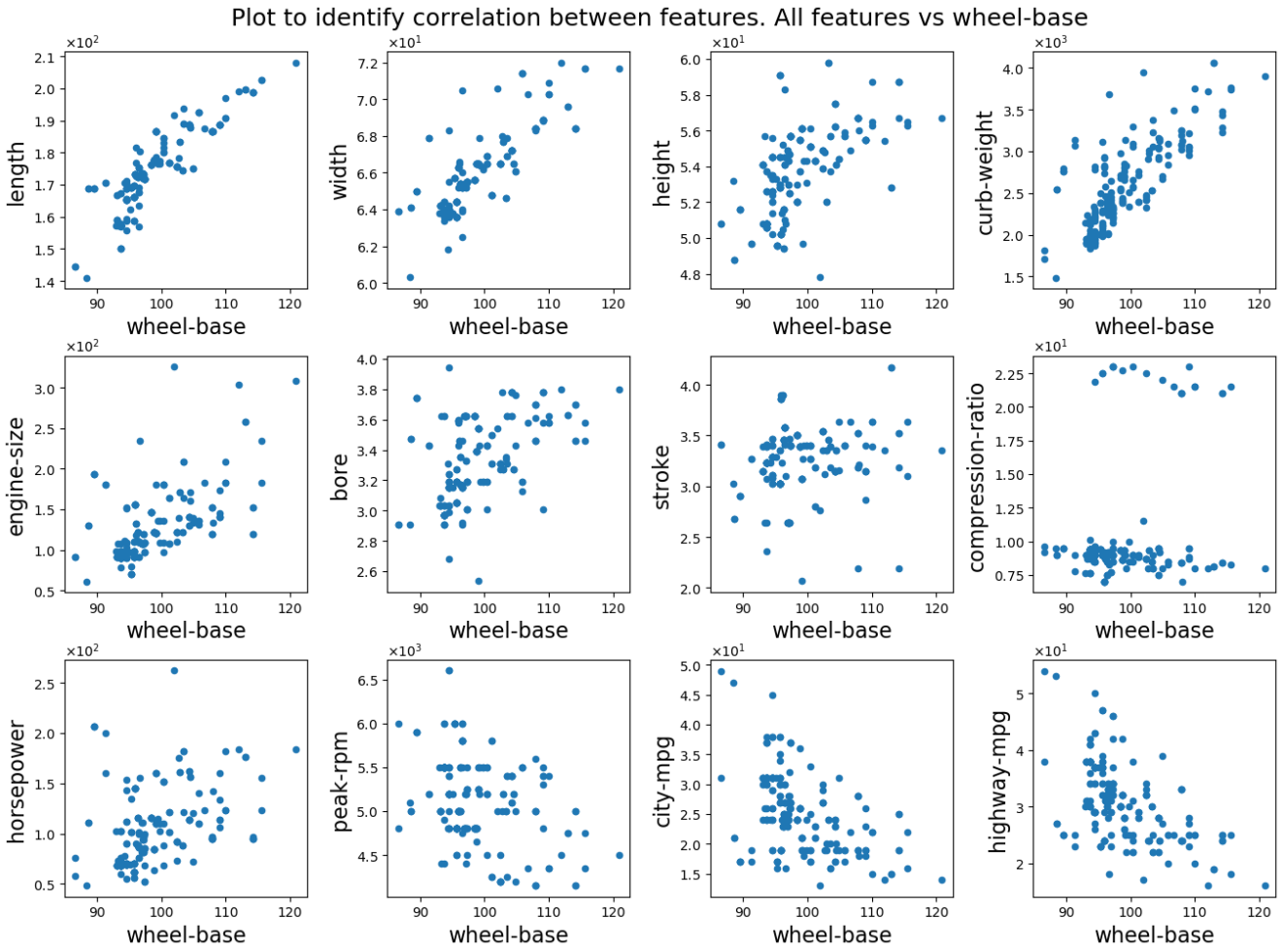
Lenght show strong correlation with wheel-base, width and curb-weight, and also a weak correlation with engine-size, horse-power and bore. Also, we can see a considerable negative correlation with city and highway-mpg.



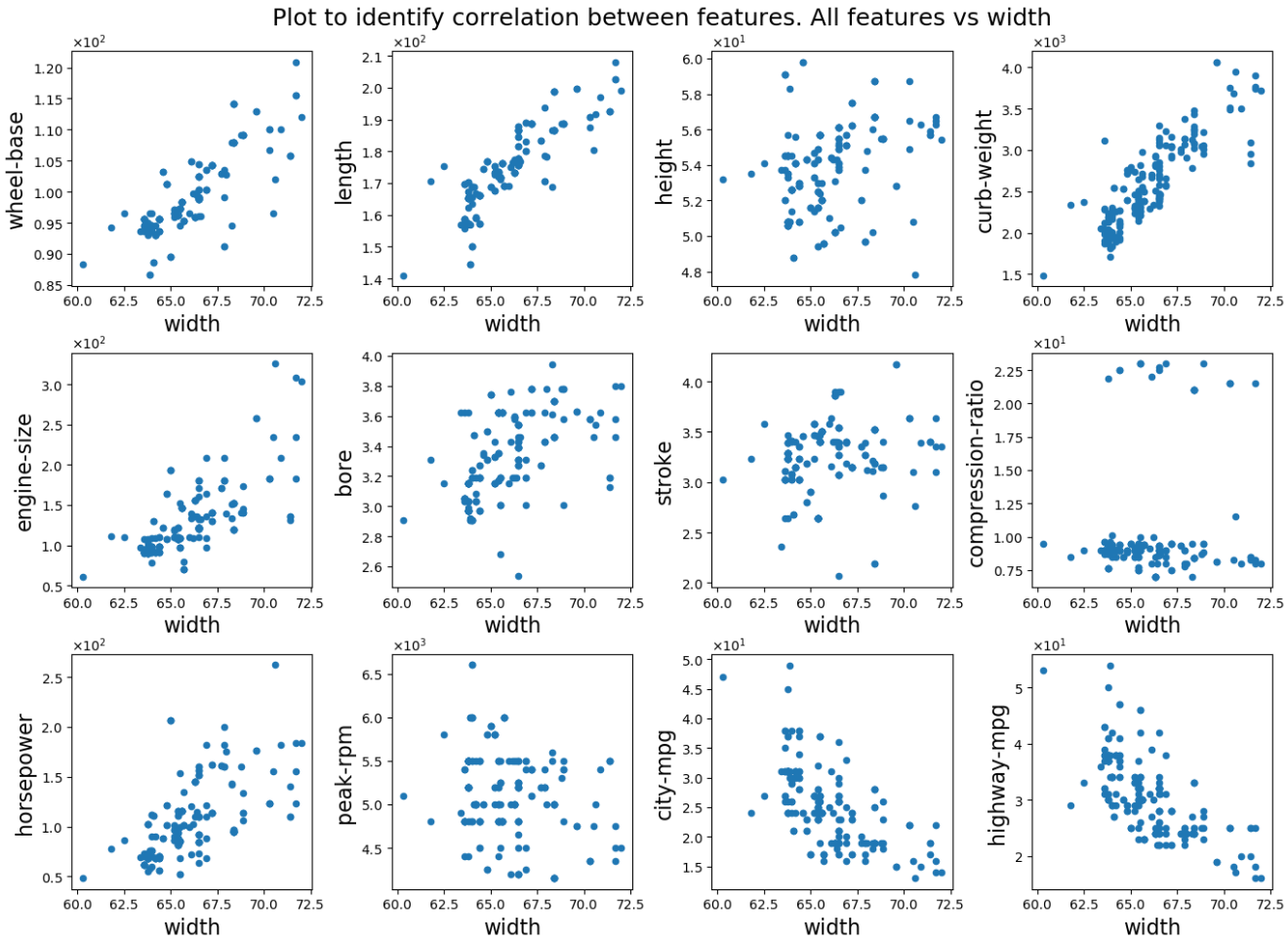
Peak-rpm doesn't seem to be correlated with any of the other features. It also didn't show any relation to the target value (recall from item c).



Stroke doesn't seem to be correlated with any of the other features. It also didn't show any relation to the target value (recall from item c).



Wheel-base has a considerable correlation with length and curb-weight, and also some weak correlation with width.



Finally, width is highly correlated with length and curb-weight, and also weakly correlated to wheel-base, engine-size and horsepower. There is also some weak negative correlation with city and highway-mpg.

Now that we have looked at all these plots, we can separate the 13 features in 4 blocks, which are very similar to the 4 blocks that we described in item c. These blocks roughly satisfy the property that within each block, features are strongly correlated, and pairs of features from different blocks are at most weakly correlated. These blocks are

B1) height, stroke, compression-ratio and peak-rpm.
B2) engine-size and horse-power
B3) wheel-base, length, width, curb-weight and bore
B4) city-mpg and highway-mpg

The only change here from the blocks of item c is that wheel-base was moved from the 1st to the 3rd block. Taking into account these correlations between features, I would not use two features of the same block in a model simultaneously. It is interesting to note how roughly B2 is associated to engine characteristics, B3 to overall car dimensions and B4 to gas consumption (except that bore in B3 is more of a engine characteristic than overall car). This time of validation is good to indicate that we are moving in the right direction.

Part 2: Regression

2 - a) Combining our intuition obtained by observing the scatter plots in the previous questions, it makes sense to propose models that have one variable from B2, one from B3 and one from B4. It also makes sense to use the inverse of features in B4 and the square of the features in B3. Looking again at the plots in item c, it seems that curb-weight is the variable in B3 with the biggest correlation with price, so we will certainly include it. For B2 and B4, it is more difficult to choose a specific feature, therefore, we will propose these three models.

Model 1:

price = w1 * engine-size + w2 * (curb-weight)^2 + w3* (1/city-mpg)

For model 1 we picked engine-size as the representative feature of B2, and city-mpg for B4 (curb-weight will be the representative variable of B3 for all models, as explained above).

Model 2:

price = w1 * horsepower + w2 * (curb-weight)^2 + w3* (1/city-mpg)

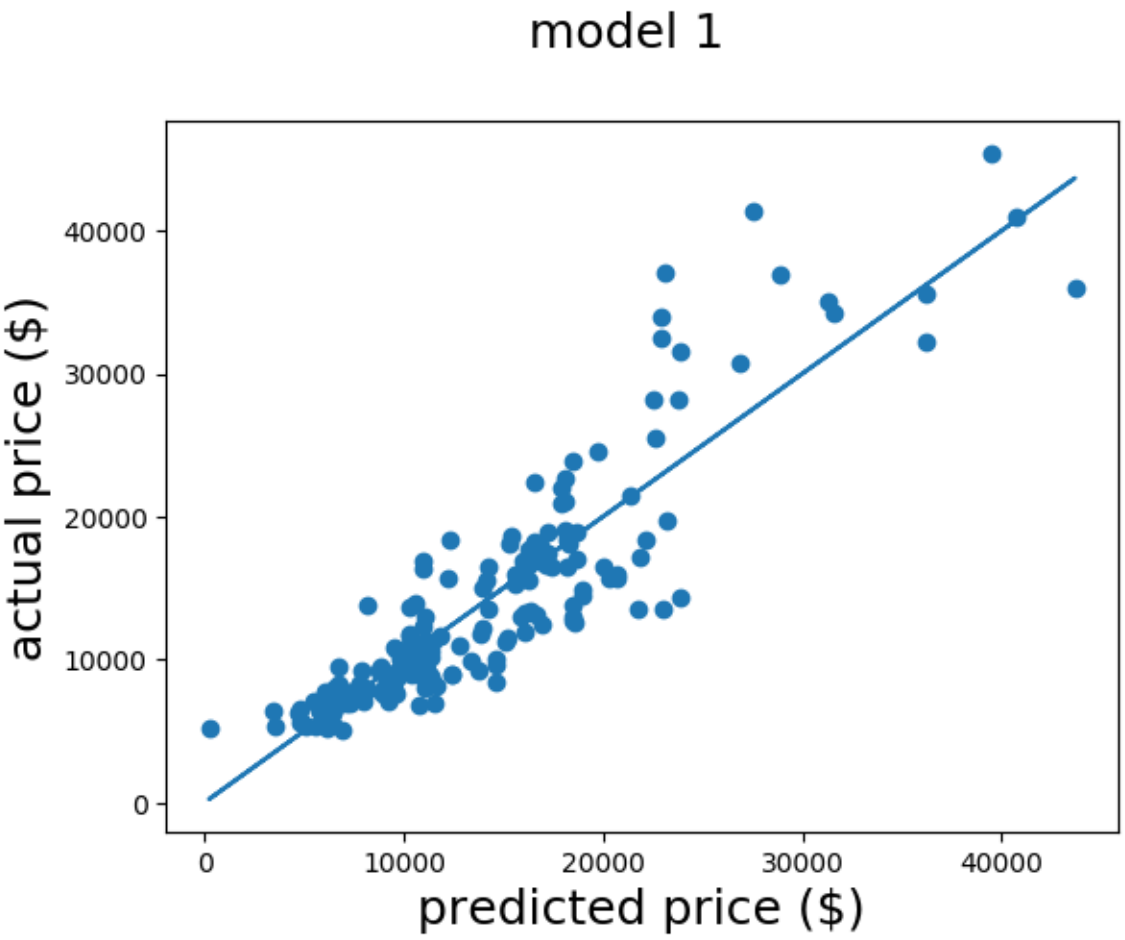
For model 2 we just changed engine-size to horsepower as the representative feature of B2.

Model 3:

price = w1 * engine-size + w2 * (curb-weight)^2 + w3* (1/highway-mpg)

For model 3, the only change from model 1 was to replace city-mpg to highway-mpg as a representative of B4.

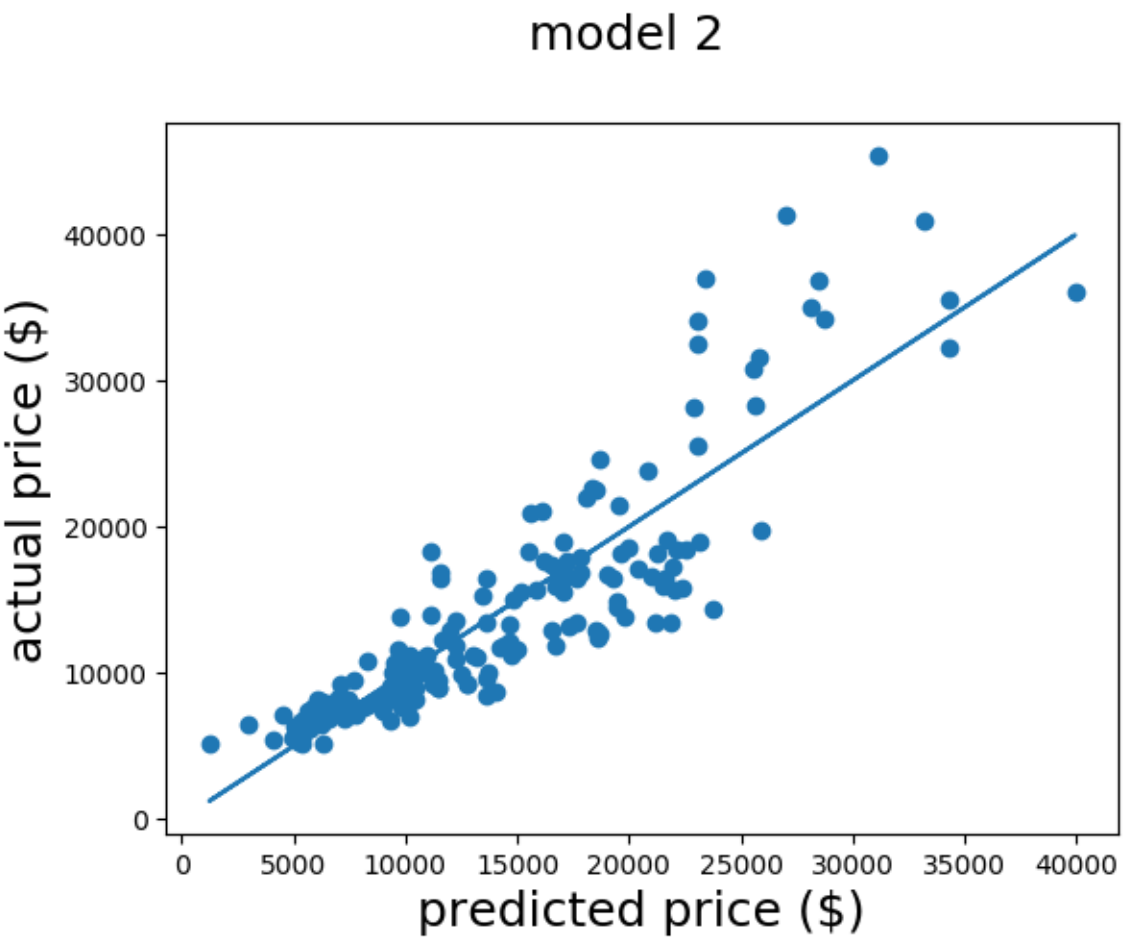
b)



price = 9.14e+01 * engine-size + 6.75e-04 * (curb-weight)^2 + 1.83e+05* (1/city-mpg)

R^2 = 0.818

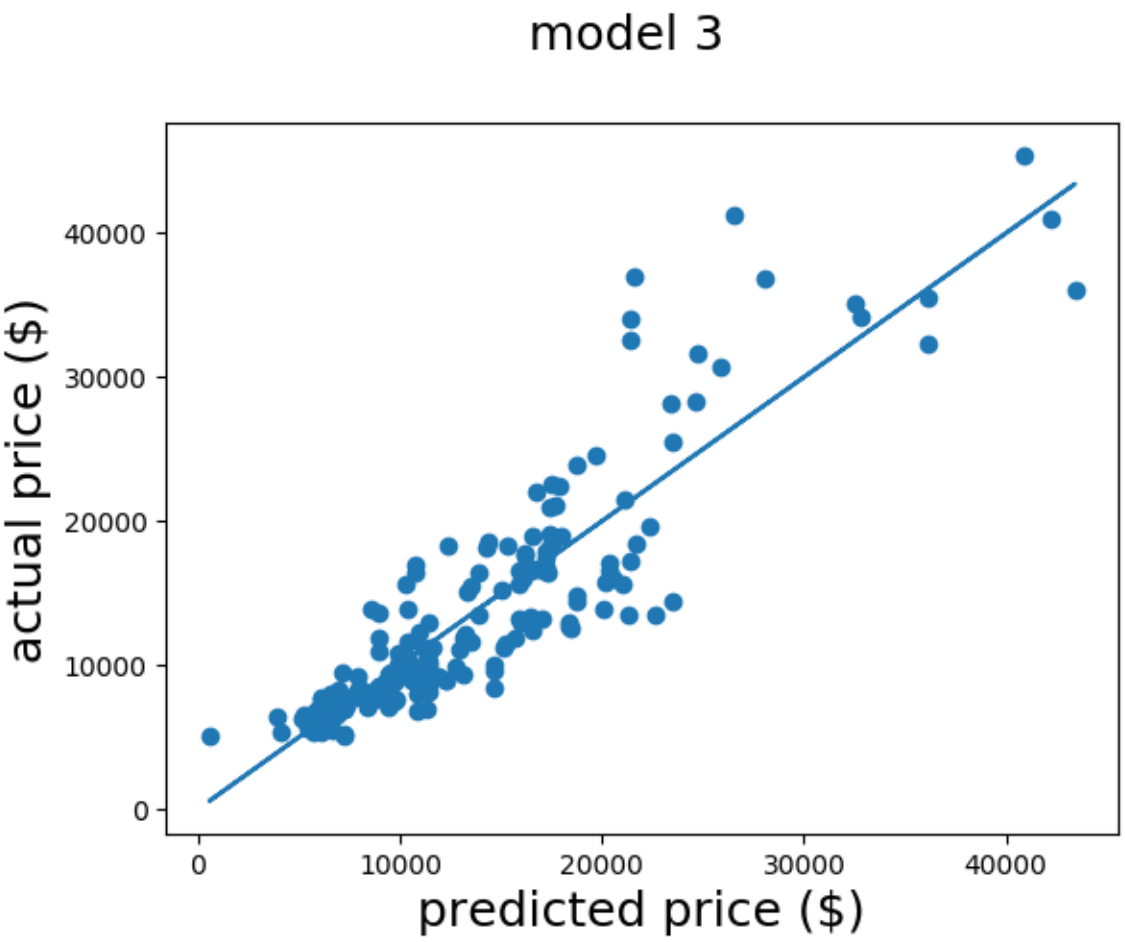
This model looks reasonably good considering the amount of data we had. There are only a few points that fall really far from the predicted price. Most of these are the more expensive cars (price > $20000, which corresponds to roughly $50000 today's dollars). In my opinion that's expected, because when we look at high-end cars, comfort and brand's reputation play a significant role in the price, and our features don't have any information about that. At lower prices, our model shows good accuracy, which indicates that these features that we used are capable of explaining most of the differences in car prices.



price = 7.55+01 * horsepower + 1.48e-03 * (curb-weight)^2 + 4.99e+04* (1/city-mpg)

R^2 = 0.790

Here, the only thing we changed from model 1 was the engine-size to horsepower. As we can see from the R^2 value and also from the plot above, that changed only increased the dispersion of the points, which means that engine-size was a better predictor for the price. We can still see here that the linear model works better for low-end cars.



price = 9.62e+01 * engine-size + 6.53e-04 * (curb-weight)^2 + 2.15e+05* (1/highway-mpg)

R^2 = 0.807

In this model, we only changed the measure of gas consumption from model 1. Now we use the highway-mpg instead of the city-mpg, and we can see a small decrease in the value of R^2. All the other characteristics of the model remain the same. Therefore, we see no advantage in making this change.

e) All the models were very similar, since we only made small changes from one to the other. Therefore, we will simply choose the one with the highest R^2, which is model 1.

3 – a) No, because r^2 measures the correlation between two series of data and R^2 is an aggregate measure of how one series of data is explained by several other series of data. To have some type of relationship between r^2 and R^2 in the multivariate setting, we would need to compute r^2 for all pairs of features and target. In the particular case when all the features are perfectly uncorrelated, it is possible to show that R^2 is the sum of all r^2.

b) R^2 is a metric that shows how the set of features together explain the target variable, whereas r^2 measures how each feature individually explains the target variable. Looking at r^2 only, we may not notice that two features explain exactly the same part of the target, so, it is unnecessary to have them in the model simultaneously.