Finding a Fuel-Efficient Car

An Exercise in Data Exploration

Greg Damico

Fall 2016

The search for an automobile with a high MPG rating is as pressing as ever. Now it is certainly true that the automobile industry has not been immune to the growing interest worldwide in alternative sources of energy. In fact, a simple internet search by the author turned up reports of cars or car prototypes being powered by air, electric battery, natural gas, biofuels, steam, nitrogen, hydrogen, ammonia, charcoal, and wood.[[1]](#footnote-1) Nevertheless, because (i) many such models are still experimental, (ii) many such models are expensive, and (iii) a change away from the current paradigm would require extensive infrastructural change, consumers will surely be using cars powered by oil (gasoline) for a long time to come.

This report, therefore, is grounded in the idea that consumers (i.e., in our case, drivers) take an interest in spending their gasoline wisely. That is to say, our main task is to discover a way of accurately predicting which car models will have high MPG ratings. But, as in many decisions facing the consumer­–especially the American consumer–there are so many options from which to choose that the task to find fuel efficiency can be overwhelming. Autos differ according to manufacturer, engine type and size, weight, and many other variables. The goal is to try to sift through all of these data to find patterns that can predict high MPG ratings.

This project is based on a (two-dimensional) dataset (available [here](http://archive.ics.uci.edu/ml/datasets/Auto+MPG) courtesy of UC Irvine) that has one row for each of 398 models of automobile. There is (of course) a column/variable for MPG rating, and other columns for various features of the car, such as number of cylinders, displacement (piston volume), horsepower, weight, acceleration, and model year.

The last of these turned out to be the criterion by which the data were ordered; it also marked a significant limitation of the dataset, since there were cars only from the 1970s and 1980s. Thus certain recent technological changes could not be accounted for in these data. Nevertheless, the range of years seemed long enough to justify us in our hope that certain trends could be discerned. Of course, though 398 is a fair number, we also could not hope to talk about every past or current automobile model. One further limitation of the dataset was that it was unknown what the odometers on these various cars read.[[2]](#footnote-2)

Our dataset was fairly clean. The first thing I did was to add names to the variables by using the ‘colnames()’ function. I then realized that there were some missing values in the horsepower column that I needed to replace with “NA”s. A harder challenge arose because of the dataset’s column ‘car name’ that lumped together the name of the manufacturer and the name of the model. Because I was interested in comparing both models of different manufacturers and models of the same manufacturer, it was important to split that column into two: I chose the new names ‘make’ and ‘model’. Because some values in the original ‘car name’ column in fact had more than two words (e.g. “buick skylark 320” in row #2), I took advantage of the ‘extra’ variable in ggplot2’s ‘separate()’ function, setting its value to “merge” so that the ‘model’ column would absorb all the words after the first. One further complication was that there were various misspellings among the manufacturer names, which led, for example, to “Chevy” being treated as different from “Chevrolet” (and both as different from “Chevroelt”!). I also merged the “Datsun” makes with the “Nissan” makes, since these are in fact names for the same company. For the sake of various visualizations, car make names were often shortened.[[3]](#footnote-3)

I also added a new column, “Country”, to record the country of origin of the various automobiles. I decided to use only five countries, categorizing the (very few) leftovers as “XX - Other”, using an ‘X’ to begin the name so as to ensure that this category would come last in an alphabetical lexicographical ordering. The five countries were: France (Peugeot and Renault), Germany (Audi, BMW, Mercedes, Opel, and Volkswagen), Japan (Honda, Mazda, Nissan, Subaru, and Toyota), Sweden (Saab and Volvo), and the USA (AMC, Buick, Cadillac, Chevrolet, Chrysler, Dodge, Ford, Mercury, Oldsmobile, Plymouth, and Pontiac).

After cleaning the data, I made a few preliminary plots: MPG vs. acceleration, MPG vs. cylinders, MPG vs. model year, and MPG vs. weight. Weight and cylinder number had a clear inverse relationship with MPG, while model year had a clear direct relationship with MPG. There was no apparent correlation between MPG and acceleration. There was perhaps a very slight correspondence between high MPG and high acceleration, but this was likely to be explained by reference to weight as a common cause of both values. Very heavy cars would tend to have both a low acceleration rate and a low MPG rating. I made a couple plots of MPG means and weight means by country.

The number of cylinders was a bit of an issue since it, though relevant, was effectively a ternary variable. That is, only 7 of the “observations”, i.e. automobiles in the dataset, had a cylinder number different from 4, 6 and 8. I was able to determine that the number of cylinders did indeed seem to have an effect on MPG, simply by using the ‘mean()’ function: Four-cylinder engines had a mean of 29.3 MPG, six-cylinders had a mean of 20.0 MPG, and eight-cylinders had a mean of 15.0 MPG. I made some box-and-whisker plots of the MPG ratings for the automobiles grouped by make, one for all of them together and three more that showed the results for four-, six- and eight-cylinder engines, respectively.

One clear result from these boxplots was that certain manufacturers had consistently high MPG ratings, despite significant variability for nearly every value of ‘make’. (Having *very little* variability here was generally the result of small sample sizes. There was only one Triumph in the dataset, for example.) In particular, Honda, Nissan, Renault, and Volkswagen all scored high ratings, while companies like AMC and Oldsmobile scored relatively low. Two important very general trends were (i) that MPG ratings tended to rise with the model year and (ii) that MPG ratings varied significantly among the models of a given manufacturer.

In fact, not only did MPG ratings vary quite a lot among models of a given manufacturer, but there was also a lot of variations among manufacturers of a given country. I made some plots to exhibit this fact.

Finally, using the ‘lm()’ function, I constructed some linear models (again, treating MPG as the dependent variable) that would help to predict high MPG ratings.[[4]](#footnote-4) My first model took in make, model year, weight, and cylinders as independent variables. The four manufacturers that stood out before stood out here again as significant: Honda, Nissan, Renault, and Volkswagen. The cylinders variable did not register as significant. Again, I chalked this up to its extremely limited range. The model had a fairly good R2 of 0.8387, while its adjusted R2 was 0.825.

Since I had already noted the effect of the number of cylinders on MPG, I threw that variable out for the next model. Model 2 used make, weight and model year as independent variables. Again the weight and model year variables registered as significant. And this model had a slightly better metric than the first: We had an R2 of 0.8387 and an adjusted R2 of 0.8255.

At this point I decided that the two key variables for predicting MPG would be weight and model year. I wanted at this point to split up my dataset into a training set and a testing set, in order to see if our linear model was indeed a good predictor of MPG. I would use the summary results from the model applied to the training set to make a prediction on the MPG we should see in the testing set. I could then compare the prediction with the actual values of MPG in the testing set.

For the sake of testing and exploratory purposes I began with a model, Model 3, that isolated only weight as an independent variable. Weight certainly still counted as significant, but our R2 was now down to a quite low 0.69. I nevertheless went ahead with what I was now calling “Experiment #1”, splitting the dataset into a training set and a testing set.[[5]](#footnote-5) I took manufacturer-means of weight from the testing set and used the results from the linear model applied to the training set to make a prediction about what I should see in the MPG column of the testing set. I then took manufacturer-means of the MPG ratings to compare them with the predictions.

In viewing the summary of the one-variable linear model applied to the training set, it was found that the cars obeyed a proportion of something like

MPG ~ -0.007677 \* weight + 46.3174.

I made a plot of the actual MPG ratings of the testing set autos as compared to the prediction, also plotting the relative errors. In fact the model’s prediction performed admirably well.

The second experiment was the real test. The procedure was the same as with Experiment #1, except that now there were two independent variables: weight and model year.[[6]](#footnote-6) Again I split my data into a training set and a testing set,[[7]](#footnote-7) and again I took means of the key variables in the testing set, so that I could compare them against the calculated predictions made on the strength of a linear model that had been applied to the training set.

The summary of this linear model suggested that MPG could be predicted according to the formula:

MPG ~ weight \* -0.006704 + (model year – 1900) \* 0.7087 - 10.45.

Again, I made two plots: the first showing the actual manufacturer-means as well as the predicted means vs. the different manufacturers, and the second showing the relative errors. The plots illustrate that this model is indeed quite accurate.

I shall conclude with some general recommendations and some suggestions for further research. The main results seem to be these

* First, weight kills: Heavier cars consistently had lower MPG ratings.
* Second, buy a foreign car: American cars consistently had lower MPG ratings. The standout manufacturers (all foreign) were Honda, Nissan, Renault, and Volkswagen.
* Third, buy a four-cylinder car: Six- and eight-cylinder cars consistently had lower MPG ratings.

There are multiple ways in which this research could be used for further work. First, because the dataset is limited by the age of the observed autos, it would be interesting to look at data on newer car models to see if any of these trends continue or if any new trends emerge. A further way to try to extend the present work would be to look for other variables that could have a significant correlation with MPG rating. It is difficult to try to conceive *a priori* of what other automotive features could play a role here, but that is of course hardly any reason to suppose that there is none. Finally, as noted at the top, the automotive industry has begun to change in favor of featuring models that use non-traditional fuels. Any move away from rapidly disappearing fuel sources is of course to be welcomed. But it is just as clear that any fuel source will present its own challenges. And in fact, no matter what source of fuel is used for automobiles, there will always be a drive to maximize that fuel’s efficiency; perhaps the results of this experiment could be useful to any such future endeavor.

1. See e.g. https://en.wikipedia.org/wiki/Alternative\_fuel\_vehicle. [↑](#footnote-ref-1)
2. That is, to the extent that MPG seems to be a function of an engine’s life, cars with fewer miles tend to be more fuel-efficient than cars with more miles. On the other hand, there seems to be some evidence that this idea about MPG declining with age is in fact a myth: See e.g. http://blog.cochran.com/wordpress/index.php/10-myths-fuel-economy-gas-mileage/. [↑](#footnote-ref-2)
3. For certain charts I used a three-letter abbreviation for each car manufacturer, as follows: AMC: “amc”; Audi: “aud”; Buick: “bui”; Cadillac: “clc”; Chevrolet: “chv”; Chrysler: “chr”; Dodge: “dod”; Fiat: “fia”; Ford: “frd”; Honda: “hda”; Mazda: “maz”; Mercedes: “mrs”; Mercury: “mrc”; Nissan: “nis”; Oldsmobile: “old”; Opel: “ope”; Peugeot: “peu”; Plymouth: “ply”; Pontiac: “pon”; Renault: “ren”; Saab: “sab”; Subaru: “sub”; Toyota: “toy”; Volkswagen: “vks”; Volvo: “vvo”. [↑](#footnote-ref-3)
4. Summaries of these models confirmed the boxplots’ findings that the identity of the manufacturer could make a big difference to the MPG rating. [↑](#footnote-ref-4)
5. Since the splitting of my data invoked a random sampling procedure, I set a seed number for the sake of reproducibility. I put 300 of the 398 into the training set and the other 98 would constitute the testing set. [↑](#footnote-ref-5)
6. The basis for this experiment was in fact, by the original numbering, Model #5 as opposed to Model #4. The latter was a poorly performing model of MPG that took only model year as an independent variable. [↑](#footnote-ref-6)
7. See previous note. [↑](#footnote-ref-7)