Capstone Project Milestone Report

1. THE PROBLEM REVISITED: It is certainly true that the automobile industry has not been immune to the growing interest worldwide in alternative sources of energy. A simple internet search by the author turned up reports of cars or car prototypes being powered by air, electric battery, wood, 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.

What is required, therefore, is 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.

1. THE DATASET: My dataset (available [here](http://archive.ics.uci.edu/ml/datasets/Auto+MPG) courtesy of UC Irvine) 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 cannot be accounted for in these data. Nevertheless, the range of years seems long enough to justify us in our hope that certain trends may be discerned. Of course, though 398 is a fair number, we also cannot hope to talk about every past or current automobile model. A further limitation is that it is unknown what the odometers on these various cars read. If nothing else, 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.[[2]](#footnote-2)

The 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 cars both of different manufacturers and as models of the same manufacturer, it was important to split up 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. A 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.

1. PRELIMINARY FINDINGS: After cleaning the data, I tried a few preliminary plots: MPG vs. cylinders, MPG vs. model year, and MPG vs. weight. The number of cylinders was going to be a bit of an issue since it 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. But by using the ‘mean()’ function, I was able to determine that the number of cylinders did indeed seem to have an effect on MPG. 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 early 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.) Honda, Nissan, Renault, and Volkswagen all scored high ratings, while companies like AMC and Oldsmobile scored relatively low. I next tried, using the ‘lm()’ function, to construct some linear models that would help to predict high MPG ratings. Summaries of these models confirmed the boxplots’ findings that the identity of the manufacturer could make a big difference to the MPG rating.

1. NEW DIRECTIONS: I was not so very surprised by what we might call the “inter-manufacturer MPG variability”, i.e. the fact that some manufacturers had much higher MPG ratings than others. But I had not been expecting that there would be so much “intra-manufacturer MPG variability”, i.e. the MPG gap between the most and the least fuel-efficient models of a single manufacturer. I decided that box-and-whisker plots were a good way of depicting this. Moreover, it was clear that the ‘make’ variable was going to play a key role once it came time to summarize my results and to draw final conclusions.

Another result from the modeling was that the weight of the automobiles was statistically a very significant factor. Upon isolating weight as the sole characteristic from which to construct a linear model, where R^2 was calculated at a not-bad 0.69, it was found that the cars obeyed a proportion of something like

MPG ~ -0.007677 \* weight + 46.3174

I plan to split up the dataset into a training set and a test set in order to see whether this is indeed a good predictor of MPG rating.

1. See e.g. https://en.wikipedia.org/wiki/Alternative\_fuel\_vehicle. [↑](#footnote-ref-1)
2. See e.g. http://blog.cochran.com/wordpress/index.php/10-myths-fuel-economy-gas-mileage/. [↑](#footnote-ref-2)