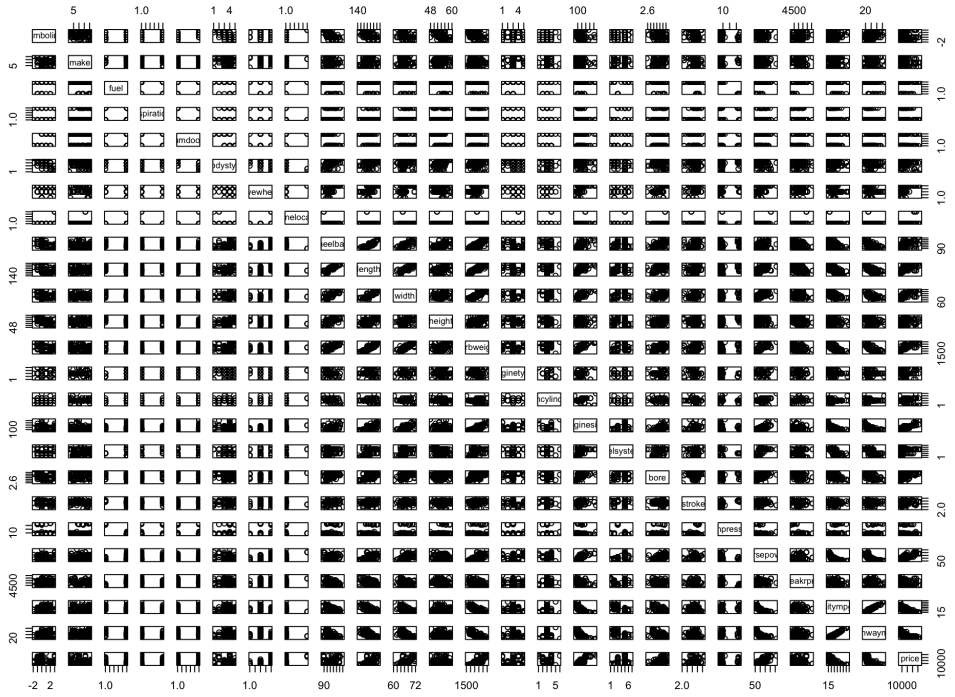
Linear Regression Project

I chose the automobile dataset for this project. Initially viewing the data set as a whole was unproductive and didn't provide clear patterns. What I did from here is try to find relationships by limiting down the size of the plots. I narrowed down the visualizations and focused on variables that appeared most strongly associated with price. I am testing whether horsepower, city mpg, highway mpg, engine size, and curb weight are significant predictors of automobile price. For each predictor, I recorded the corresponding R^2 and AIC values.

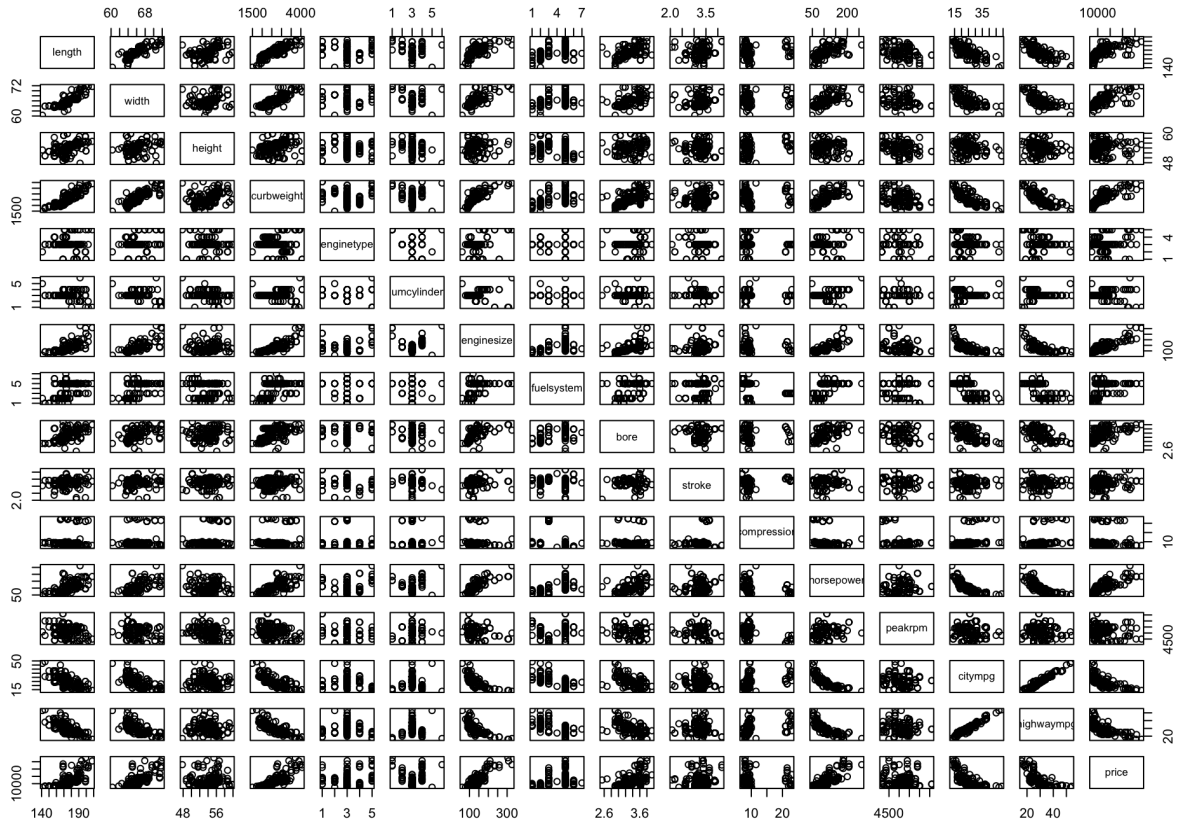
Whole Auto Dataset



When I plotted the whole dataset at once, the graph looked extremely messy. All the variables overlapped and it was hard to see any clear patterns at all. This showed me that looking at everything together wasn't helpful because too many relationships were being mixed together.

So I realized I needed to break the data into smaller pieces to better understand how price relates to other features.

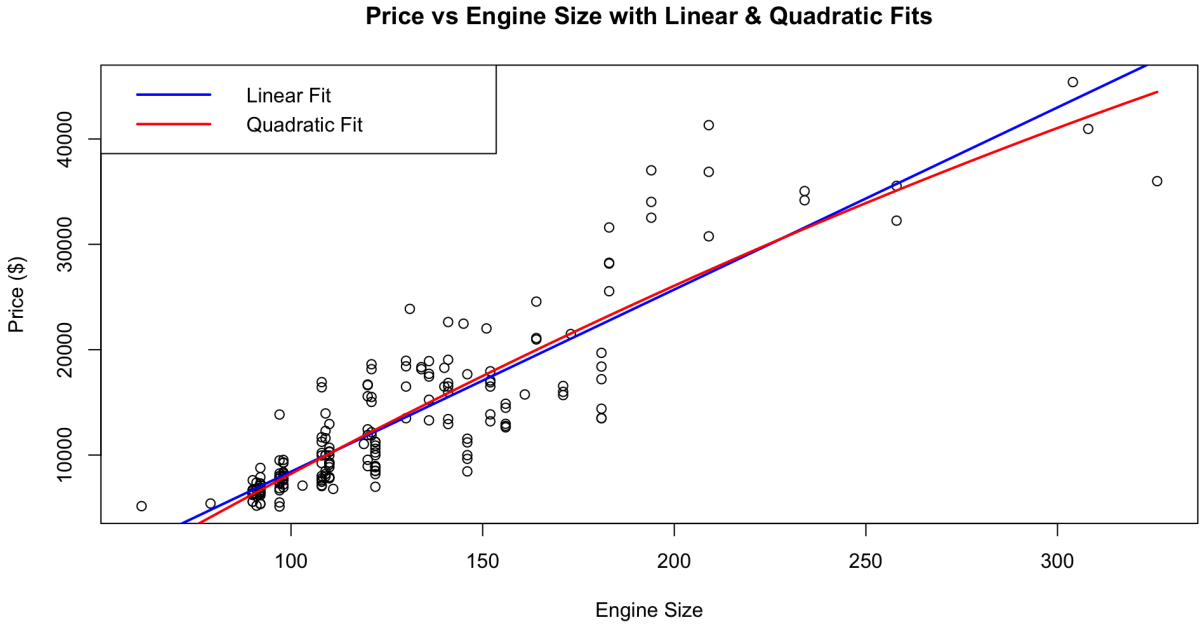
Smaller Plot



After zooming in on a smaller set of variables, I could finally start noticing patterns. Price looked like it might increase as engine size, horsepower, and curb weight increased. On the other hand, cars with better city and highway MPG seemed cheaper. These smaller plots helped me pick out which predictors actually looked useful for modeling price. That's why I decided to test those five variables in the regression models.

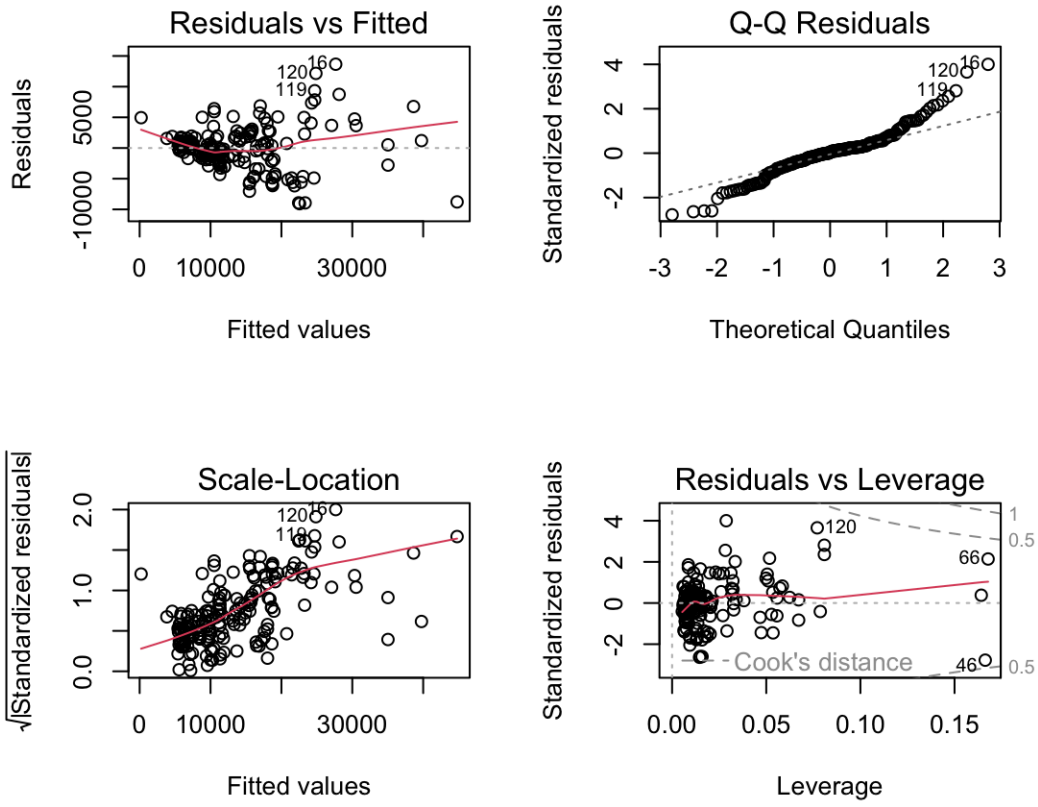
Predictor	R^2	AIC
Price ~ horsepower	0.6583	3817.784
Price ~ city mpg	0.4967	3892.529
Price ~ highway mpg	0.5147	3885.500
Price ~ engine size	0.7888	3724.902
Price ~ curb weight	0.6963	3795.058

To find the best single predictor of automobile price, I fit five simple linear regression models, each using one predictor: horsepower, city MPG, highway MPG, engine size, and curb weight. I compared the models using R^2 and AIC. Among all one-predictor models, engine size performed best ($R^2 = 0.7888$, AIC = 3724.902). Therefore, I selected price ~ engine size as Model 1 (m1). A scatterplot with the fitted linear and quadratic regression line is shown below.



For Model 1, the scatterplot shows a clear trend: cars with bigger engines usually cost more. The straight-line (linear) model fits this relationship really well, with an R^2 of 0.7888, meaning engine size alone explains almost 79% of the differences in car prices. I also tested a quadratic model to check if the relationship curves for cars with very large engines. The quadratic line does bend slightly at the high end, which suggests that prices don't increase quite as fast once engines get extremely big. However, the quadratic model only improves the AIC by less than 1 (3723.989 vs 3724.902), which is basically nothing. Since the improvement is tiny and the linear model is simpler and still performs great, the linear model is the better overall choice for Model 1.

Model 2 Diagnostic Plots



To check how well Model 2 works, I looked at the four diagnostic plots. In the Residuals vs. Fitted plot, the errors are small for cheaper cars but spread out a lot more for higher-priced ones. This shows heteroscedasticity, meaning the model isn't as accurate for expensive cars. There's also a slight curved pattern, which suggests the model doesn't fully capture the true relationship between the variables and price.

In the Normal Q-Q plot, the points mostly follow the straight line in the middle, which is good, but the tails bend away quite a bit. This means the residuals aren't perfectly normally distributed and there are some noticeable outliers. Mainly luxury cars with unusually high prices.

The Scale-Location plot tells a similar story, the residuals get more spread out as fitted values increase. This again shows that prediction error rises for high-end vehicles, so the model fits average cars better than the very expensive ones.

Finally, the Residuals vs. Leverage plot shows a few influential observations that pull on the model more than the rest. These points likely represent rare or unusual cars that don't follow the typical pricing patterns in the dataset.

Model	R^2	AIC
Model 1 (engine size)	0.7888	3724.902
Model 2 (price ~ enginesize + curbweight + horsepower)	0.8166	3699.624

Conclusion

When comparing the two models, Model 2 does a better job overall. Model 1, which only uses engine size, already explains a lot of the price variation ($R^2 = 0.7888$) and has an AIC of 3724.902, so it's a strong model by itself. But when I added curb weight and horsepower in Model 2, the R^2 increased to 0.8166, meaning it explains even more of the differences in car prices. Also, the AIC dropped to 3699.624, which tells us Model 2 gives a better fit while still being efficient. So, even though Model 1 is simple and works well, Model 2 clearly explains more variability and should do a better job predicting future car prices.