Final Report

Project Team 6
Jack Beck, Dria Fabrizio, Jordan Waldroop

Executive Summary

In order to align with the company's goal to remain environmentally friendly when purchasing fleet vehicles, an analysis was conducted to determine the most fuel efficient vehicles. Under the assumption that this is based in the 1980's, with ample availability of vehicles in the dataset, and the price point is not relevant to the scenario, we have determined the specifications for the ideal fleet vehicles for the company.

The key indicator variables are horsepower, weight, and cylinder count. The recommended fleet vehicles will meet the following specifications: 75-150 horsepower, 2000-2500 pounds, and 4 cylinders.

When the recommended specifications were applied to the cars available in the dataset, we found that the 1980 Volkswagen Rabbit (41.5 MPG, 76 HP) and the 1980 Datsun 510 hatchback (37.0 MPG, 92 HP) were the best choices for fleet purchase. Overall, there are 60 vehicles in the dataset that meet the specifications for the ideal fleet vehicle.

Business Understanding

Our project aimed to mimic analysis that might be done by a company looking to purchase a fleet of cars for employee use. Our theoretical company is environmentally conscious and would like to ensure that the cars selected for the fleet are fuel efficient, which will help to benefit the environment while also keeping fuel costs as low as possible. When choosing data, we prioritized datasets that would best fit the needs of our analysis. As a result, we choose a data set that features cars from the years 1970-1982. Because of this, our theoretical company is assumed to be operating in the early 1980's.

By analyzing the available data on different vehicles, we were able to identify the class of vehicles that would be best suited for meeting the company's needs. The data provided *'Car Name'*, *MPG* (the average miles per gallon), *Cylinders*, *Displacement*, *Horsepower*, *Weight*, *Acceleration*, *Model_Year*, and *Origin* (country of origin). Using this information, the team performed a multiple linear regression to determine which of the features in the dataset were the best indicators of fuel efficient cars. Using the results from the regression, the company could determine which features to look at, hence narrowing its search and more efficiently finding cars that suit the company's needs.

Data Understanding and Preparation

The dataset used in the project was an automobile miles-per-gallon (*MPG*) dataset, which featured a number of variables that were potentially related to *MPG*, the variable chosen as the dependent variable for this analysis. Based on the variables in the dataset, our team determined it was the ideal choice for solving our business problem. This dataset was obtained from the UCI Machine Learning Repository. This repository contains 559 data sets, which UCI maintains as a service to the machine learning community. The variables contained in this dataset are as follows:

- *Mpg* (continuous variable)
- *Cylinders* (multi-valued discrete variable)
- *Displacement* (continuous variable)
- *Horsepower* (continuous variable)
- *Weight* (continuous variable)
- *Acceleration* (continuous variable)
- *Model_Year* (multi-valued discrete variable)
- *Origin* (multi-valued discrete variable)
- *'Car Name'* (string variable)

The dataset that was used was already relatively clean. The only actions necessary to work more easily with the data available were to change the header fields to be more easily recognizable, as well as replacing "?" values with null values from the *Horsepower* column and then converting the remaining values from characters to numeric. Otherwise, it was not necessary to do any other data-cleaning to perform our analyses.

Modeling

Before performing any analyses, we first had to look at our business problem and determine what our dependent variable would be when performing the analyses. As the business problem is centered around finding the most fuel efficient vehicles, the variable *MPG* was chosen as the dependent (x) variable, meaning that our regression analysis aimed to predict *MPG* using the chosen features.

We searched both forward and backward to evaluate the variables' metrics in regards to our regression models. We first evaluated each individual variable against *MPG*, determining the statistical significance of each based on the p-value, R^2, and the adjusted R^2 outputs. The p-values were used to determine the statistical significance of each predictor variable, while the R^2 and adjusted R^2 were used to determine the proportion of the variance in the dependent variable that is explained by the independent variables.

Based on the preliminary multiple linear-regression models created, our team found that the variables: *Displacement*, *Acceleration*, *Model_Year*, and *Origin*, were not statistically significant, while the variables: *Cylinders*, *Horsepower*, and *Weight* were statistically significant. The models with the highest adjusted R^2 were those that contained the previously listed statistically significant variables, suggesting that these variables should be the ones focused on in our subsequent models. It is worth noting that the *Cylinder* variable was treated as a factor in order to break it into each respective categorical variable, each representing a different cylinder count (i.e. 3,4,5,6,8).

In order to confirm that the variables that did not have any statistical significance in regards to the dependent variable, we did a forward search of the variables, then removed each variable until we again confirmed which predictor variables were in fact statistically significant. In the forward search, there was no change in our conclusion about the statistical significance of each of the predictor variables.

While performing both the forward and backward searches, we also plotted the results of the regression models to scatterplots, to help visualize any relationship between the variable(s). After reaching a conclusion regarding the statistically significant variables, we then measured the VIF for the models that appeared to be most relevant, to test for collinearity between the chosen variables. One of the initial models, *mod4*, contained the statistically insignificant variables displacement and acceleration. The VIF calculated for this model had several VIF's close to or greater than 10, suggesting that there was collinearity between variables in the model. We suspected that this collinearity was between *Horsepower, Weight, and Displacement*, since from a logical perspective, *Horsepower* should increase as *Displacement* increases and as *Weight* decreases. Based on their high p-Values and suspected collinearity, *Acceleration* and *Displacement* were removed from the final models.

Analysis of the various plots created revealed that while our models did imply a linear relationship, the fit was not as close as we would have liked (*SEE FIG 1*). The scatter plot of *MPG, Horsepower,* and *Weight* had a curve to it that looked to be of an exponential nature. In order to account for this in a new linear model, our team decided it would be best to run the same regression, but with all variables converted to log scale. This was done by taking the log() of each variable and appending it to the data table as its own new column (i.e. logMPG). After creating a log variable for each of the statistically significant variables, new scatter plots were created to test the effectiveness of this transformation. Our first log plot, which plotted *logMPG, logHorsepower, and logWeight* provided encouraging results, with the scatter plot following an obvious linear relationship (*SEE FIG 2*). From here we created a second log plot, which added the categorical variable *cylinders* to the plot in the form of the shape of each point. Note that the *Cylinders* feature was not logged as it is a multi-valued discrete variable. The results of this second plot reaffirmed our assumptions that 4 cylinder cars with low weight and low-moderate horsepower were the most fuel efficient cars (*SEE FIG 3*).

The encouraging results from the scatter plots were further supported by the statistics provided by the multiple linear regression models created using the new log variables. The first log model focused only on *logHorsepower and logWeight* as independent variables. This model produced

an adjusted R^2 of 0.7949, with each of the features having a statistically significant p-value. The VIF was also calculated on each variable to test multicollinearity, with each variable having a VIF of ~ 4, suggesting that there was no significant collinearity. However, based on our earlier scatter plots and non log models, we thought that the *cylinders* variable would help to explain additional variance. This ultimately led to what would be our final model, which predicted *logMPG* using *logHorsepower, logWeight,* and *Cylinders* (categorical). This new model had an adjusted R^2 of 0.8076, with most variables having statistically significant p-values. However, it is worth noting that the variables for 6-cylinder and 8-cylinder engines had statistically insignificant p-values. We decided to keep these variables in the model because of the promising results of the second log scatter plot.

From this point it was assumed that 6 and 8 cylinder engines are objectively less fuel efficient than 4 cylinder engines. Once the final model was created, we once again calculated the VIF for each variable to test for multicollinearity. Our initial VIF results suggested that there was collinearity between the various *Cylinders* variables, with each *Cylinders* variable having a VIF > 15 (other than the 5 cylinder, which only has 3 data points). However, we assumed that this high VIF came from collinearity between the other *Cylinder* categories. A new model was created to test this, with *Cylinders* being treated as one variable. This produced acceptable VIF results.

The relevant statistics of the final model can be found by running the summary() function on the model, *logmod8*, which is located on line 154 of the R-script. From this point, the team began plugging various numbers into the final regression equation to test the models predicting capabilities, which ultimately allowed up to answer our business questions. The final regression equation is as follows:

$$logMPG = 8.59 + (-0.313)x1 + (-0.532)x2 + (0.238)x3 + (0.337)x4 + (0.119)x5 + (0.099)x6$$

$$Where\ x3,\ x4,\ x5,\ x6 = 0\ OR\ 1$$

Conclusion and Discussion

Using the final regression equation above, our team tested a number of different indicator variable values to determine the best specifications for high MPG cars. It is worth noting that when plugging in the variables, the final result must be exponentiated using base *e* to transform the result from log scale to a linear scale. Using this regression equation, we deduced that the most fuel efficient cars have a *Horsepower* of 75-150HP, have a *Weight* of 2000-2500lbs, and have 4 *Cylinders*. Using values in the middle of these ranges, 100HP, 2300lbs, and 4 cylinders, our regression equation produced an estimated MPG of 26.366, making cars of these specifications ideal for our business's use.

From here, we indexed the data set using the specification ranges listed above. Of the cars in this dataset, 60 of them met our desired specifications, with a max MPG of 41.5 and a minimum MPG of 19.0. The cars with the top 2 highest MPG rating were the VW Rabbit and the Datsun 510 Hatchback, with MPG values of 41.5 and 37.0 respectively. While these cars make ideal

candidates from a MPG perspective, many others in the list of 60 cars would be suitable choices, like the 2002 BMW, which has an impressive 113 HP while still maintaining an MPG of approximately 26.0. The entire list of indexed cars is included in the teams .R script, the majority of which will suit our business's needs. We recommend management either purchase cars that are directly on this list, or ensure that they choose cars that meet our desired specifications if they are not on the list.

One thing that could be problematic when deploying these results is that there may not be enough supply of any specific model to purchase enough cars for the entire company. As a result, the company may have to compromise and buy a handful of different models.

Furthermore, our analysis does not take into account the cost of each of these vehicles. Further research will need to be done to determine which models are the most fuel efficient and cost effective.

It is important to remember that all the suggestions in this report are only true given that our prior assumptions are true. New developments in engine technology, like increased fuel efficiency for 6-cylinder engines, could nullify our prior assumptions, which will require a new analysis to be done. Additionally, the dataset used for this analysis features 398 observations of different models of car. While this is large enough for our current analysis, having a larger data set would likely make our results more accurate if we were to revisit this in the future.

Figures/Plots

Figure 1

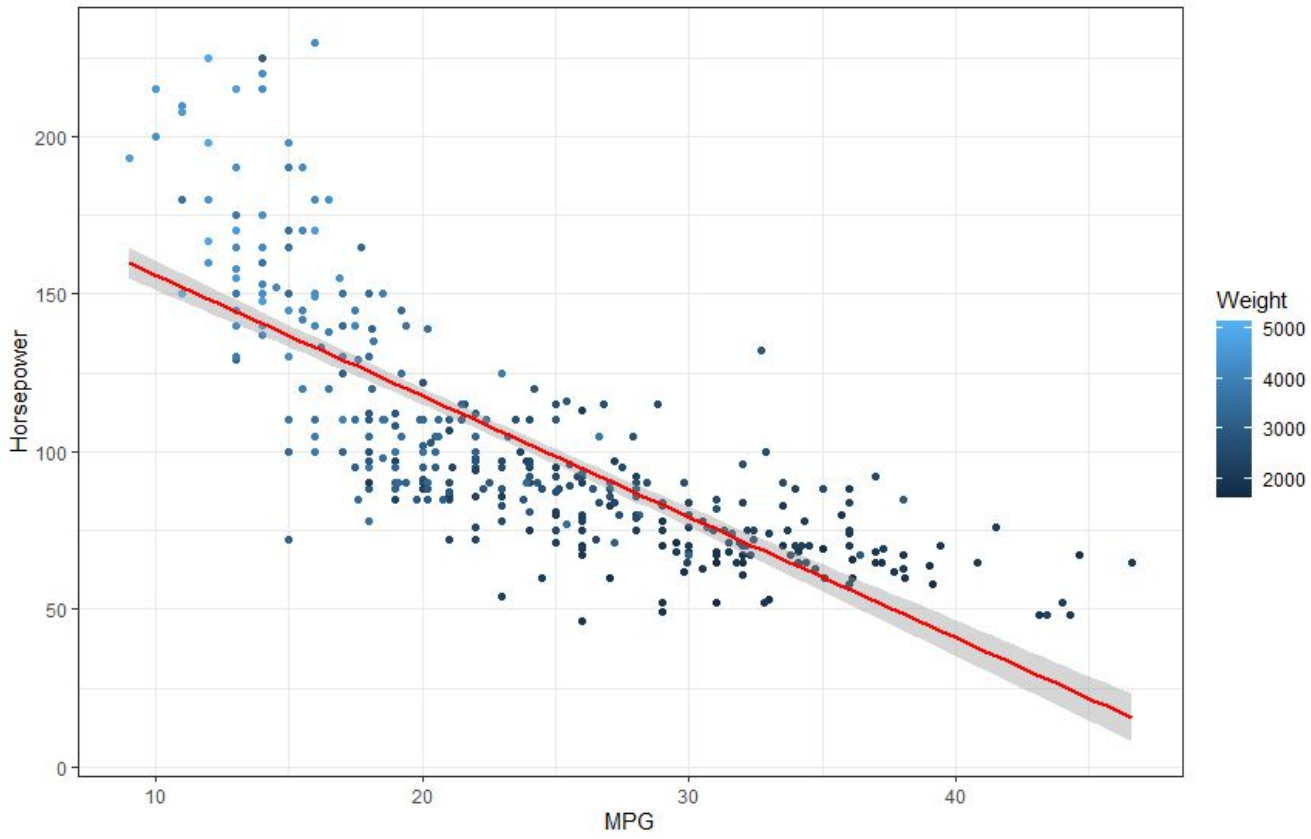


Figure 2

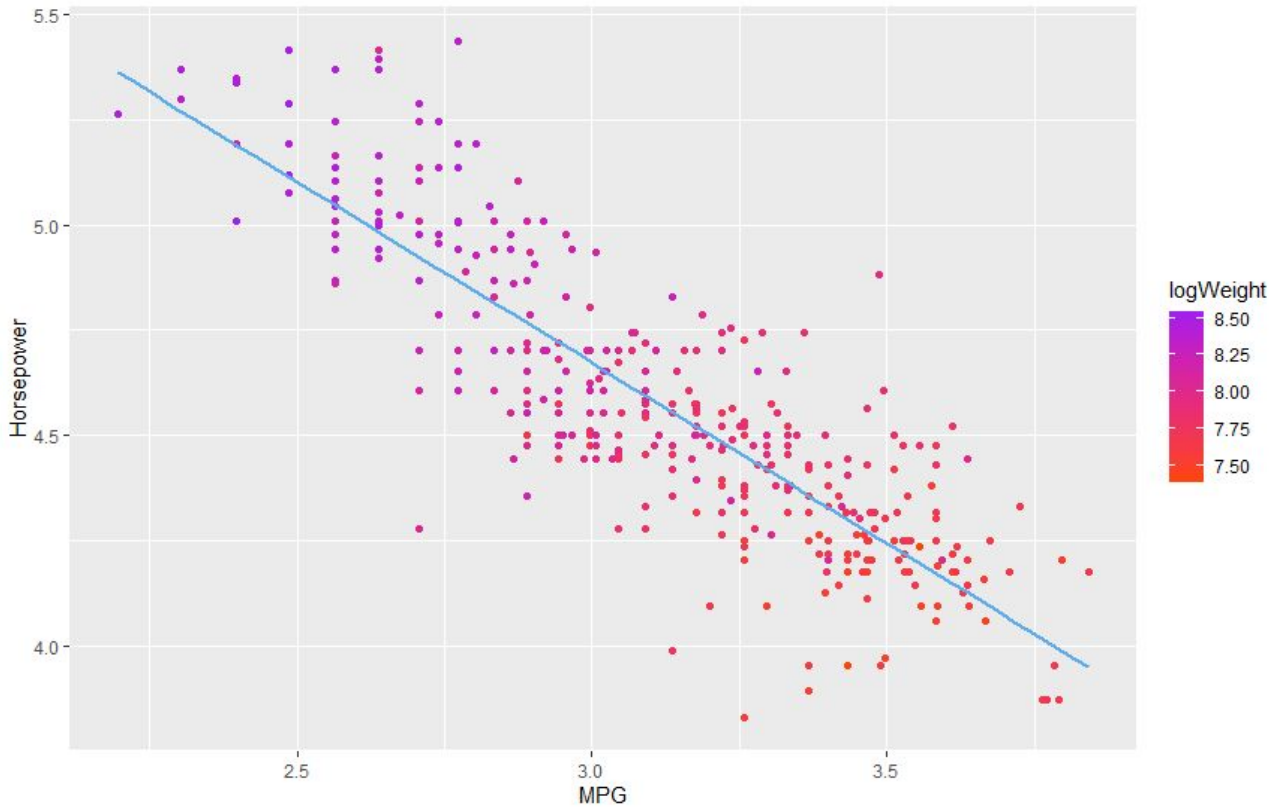
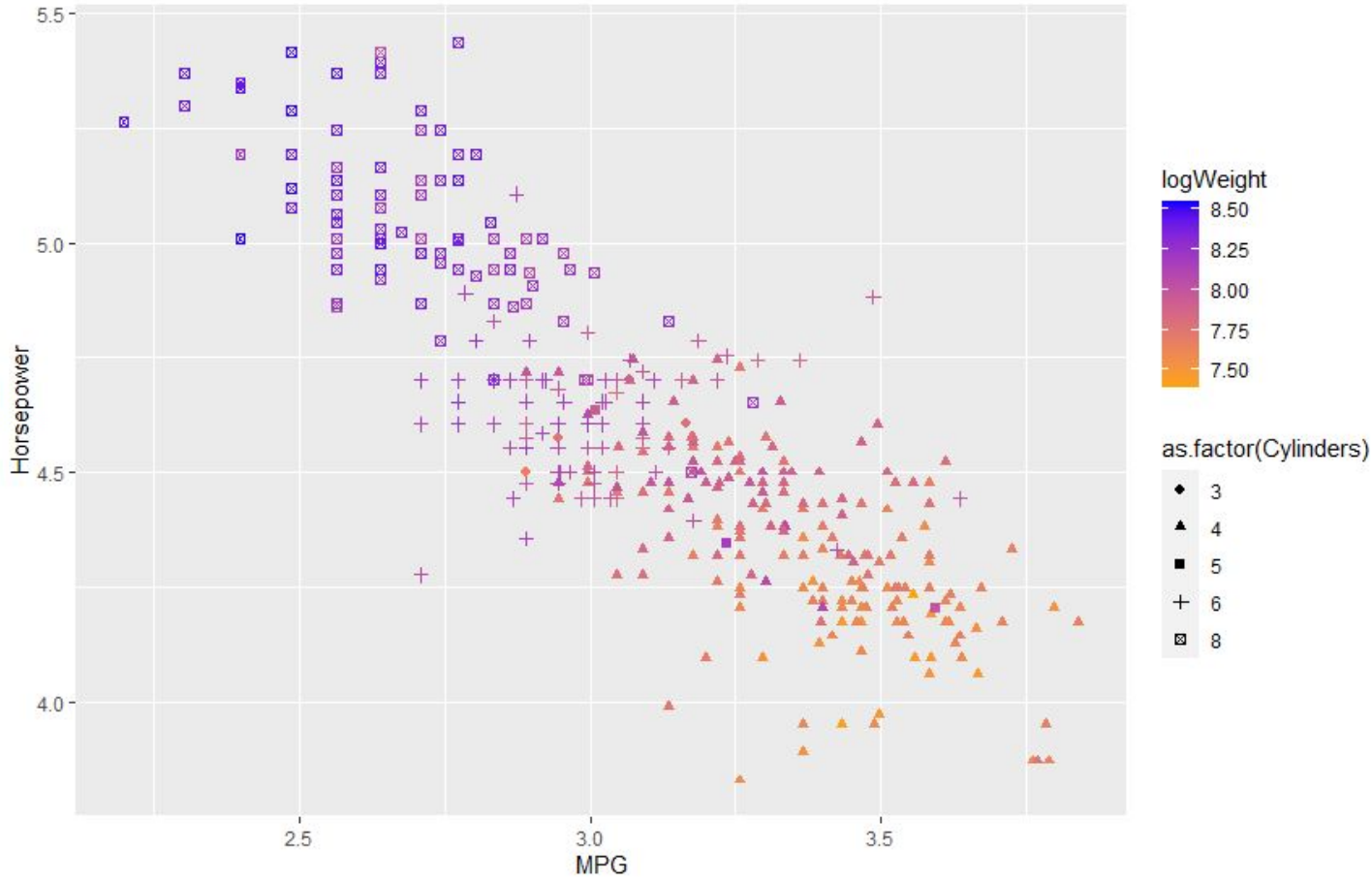

5.5
5.0
4.5
4.0
Horsepower
2.5
3.0
3.5
MPG
logWeight
8.50
8.25
8.00
7.75
7.50

Figure 3


5.5
5.0
4.5
4.0
Horsepower
2.5
3.0
3.5
MPG
logWeight
8.50
8.25
8.00
7.75
7.50
as.factor(Cylinders)
3
4
5
6
8