Karina Santoso STATS 101A

22 March 2021 Professor Almohalwas

Predicting Car Prices in the US Market

Abstract

This project explores what variables are most significant in predicting the price of an automobile in the American market and what the most accurate and effective way to use these variables to predict the price of a given car is. A multiple linear regression model was used to achieve this, which was trained with a dataset with observations of car prices and other information about the car's various other characteristics, and takes in information about a car's characteristic and outputs a numerical value representing the car's price. The goal of the project was to increase the accuracy rate of this multiple linear regression model as much as possible while keeping model complexity low.

In the Kaggle competition, my Kaggle team name was Karina Santoso Lecture 1. The R-Squared value of my training model 0.9305 and the Adjusted R-Squared value of the training model was 0.929. The R-Squared of the testing model, or the R-Squared value reported on the Kaggle leaderboard, was 0.95685, and my Kaggle final rank was 44. The total number of predictors of the final model was 5, and the total number of betas including $\beta 0$ was 23. Using the R command "UseExtractAIC(model,k=log(n))," it was found that the BIC score of the multiple linear regression model was 253.3727.

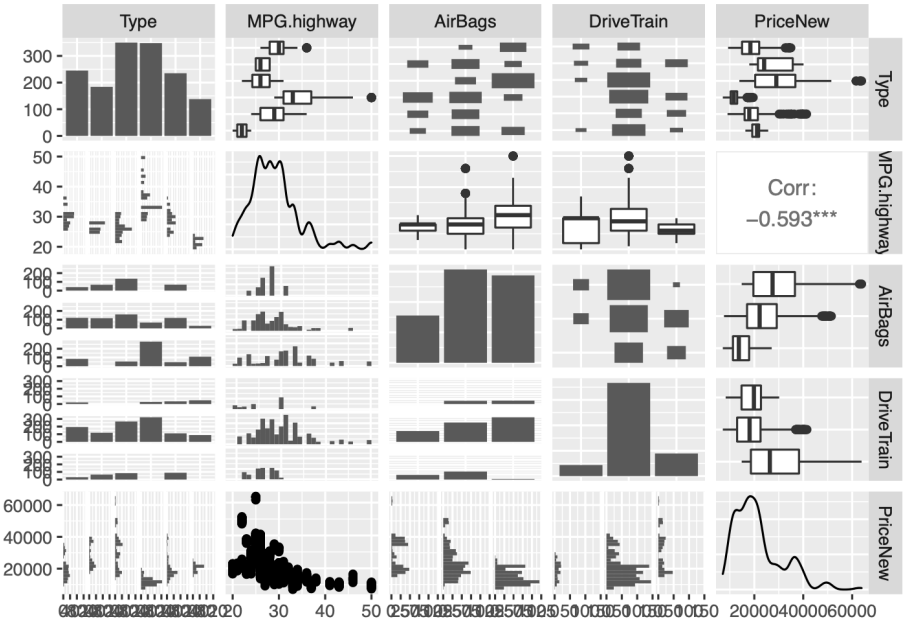
Introduction

The goal of this project was to create a model that accurately captures the most significant variables in determining a car's price in the American market to aid a Chinese automobile company, Geely Auto, to enter the US market. For a new foreign automobile company to enter the US market, it is essential to have a good understanding of the most important factors that affect the pricing of cars in the market, so that the company can price, produce, and sell their products accordingly.

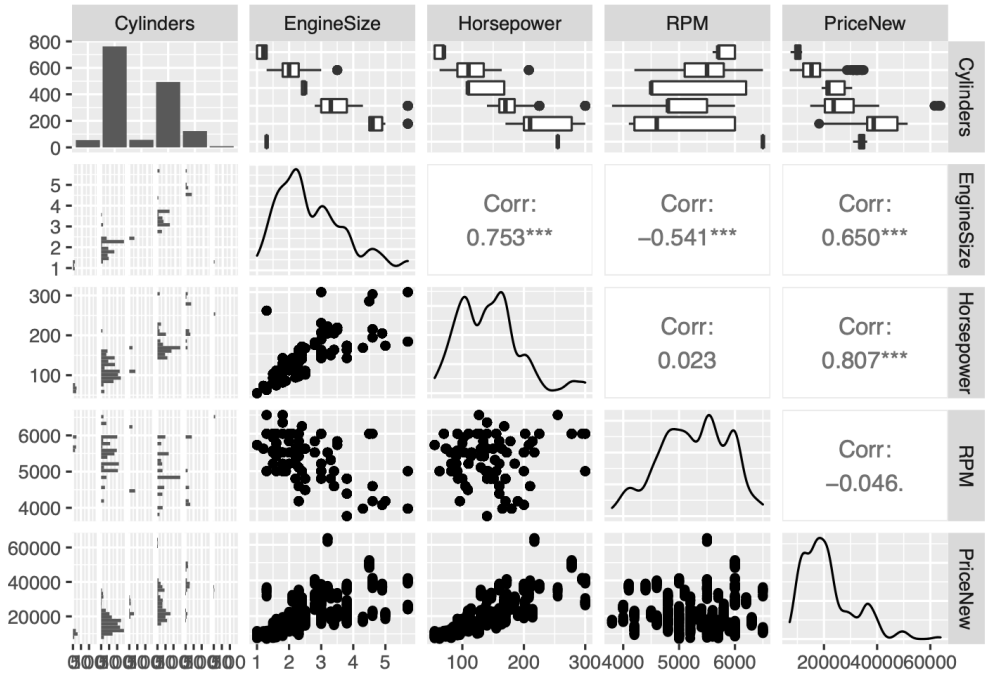
To create the model, a training dataset of 1500 observations with 24 columns of information was used. These columns were made up of a combination of both numerical and categorical data. The multiple regression model's predictor variable is the numerical variable PriceNew, which represents the price of the observed car in dollars. The goal of this project was to create a multiple linear regression model to predict the price of a car as accurately as possible given the other 23 columns of information. The other 23 variables given were: Manufacturer, Model, Type, MPG.highway, Type, MPG.highway, AirBags, DriveTrain, Cylinders, EngineSize, Horsepower, RPM, Rev.per.mile, Man.trans.avail, Fuel.tank.capacity, Passengers, Length, Wheelbase, Width, Turn.circle, Rear.seat.room, Luggage.room, Weight, Origin, and Make. Since the data was not collected over time as all columns of an observation should be filled out at the same time, there is no problem with independence in the data.

Methodology

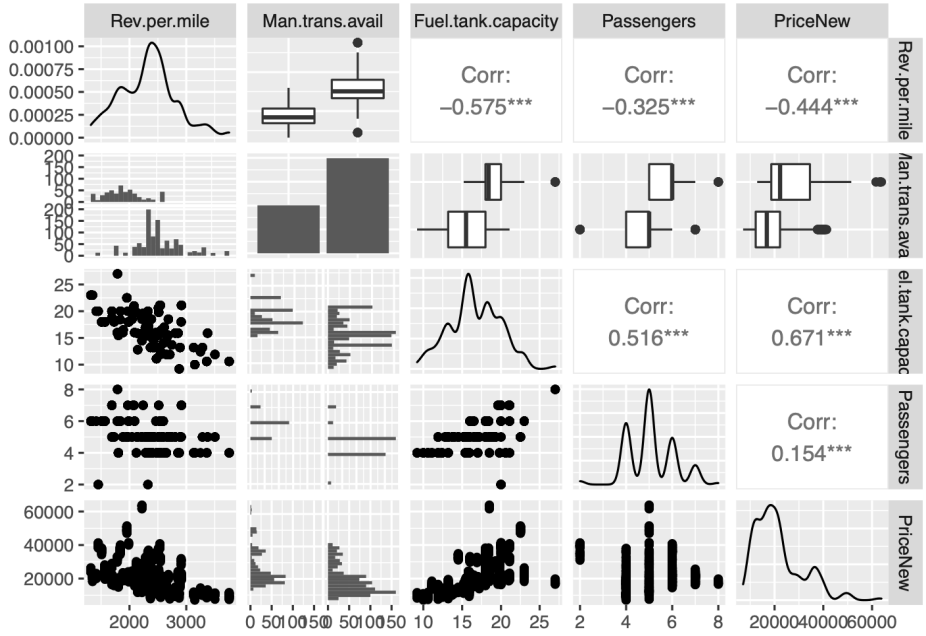
To begin exploring which of these would be the most effective in predicting the price of a car, matrix and correlation plots were utilized to visualize the relationships between the PriceNew variable and the various predictor variables. This provided a better sense of which variables have significant price changes between categories of a categorical variable or as a numerical variable increased or decreased. The Manufacturer and Make predictor variables were omitted from this analysis as they were categorical variables with too many factors. Using these variables in the multiple linear regression themselves as they were given would result in too many β coefficients and an unnecessarily complex model.



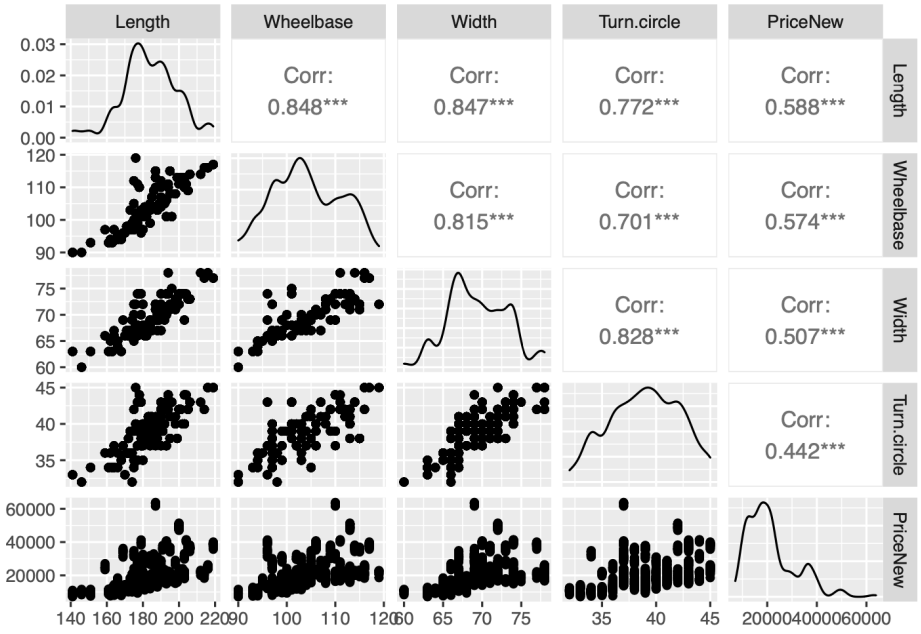
Looking at the rightmost column, it seems that the Type, MPG.highway, and possibly AirBags variables are significant, while the DriveTrain variable seems less so.



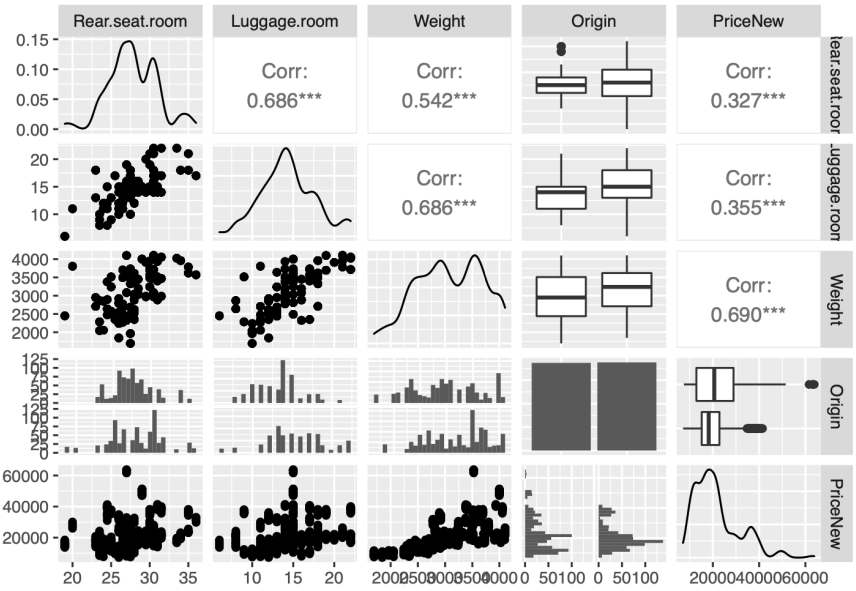
The Cylinders, EngineSize, and Horsepower variables seem to be significant, while RPM seems less so.



Rev.per.mile and Fuel.tank.capacity seem to be possibly significant variables, while Passengers and Man.trans.avail are likely not as much so.



The Length, Wheelbase, Width, and Turn.circle variables all seem to be somewhat similarly distributed and all have a medium significance in predicting PriceNew.



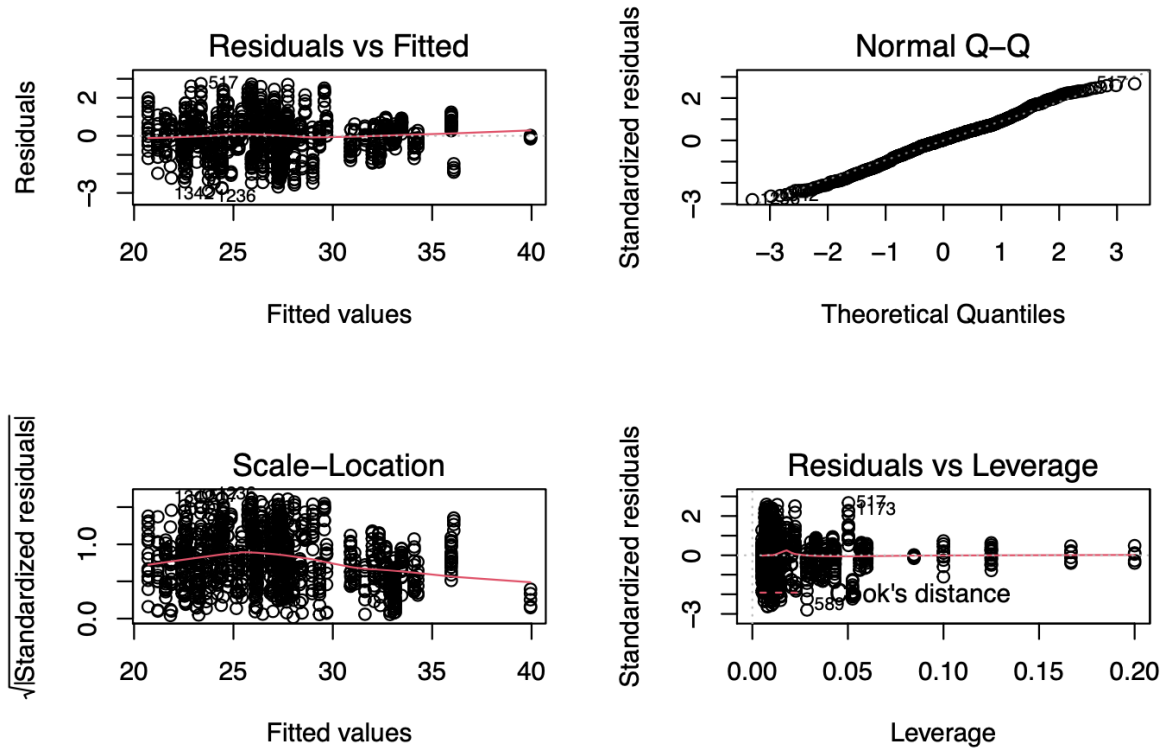
The Weight variable seems significant, while the Origin, Luggage.room, and Rear.seat.room variables do not seem to be.

```
vif(m0)
```

```
##                                    GVIF Df GVIF^(1/(2*Df))
## Type                           17.67408  5        1.332704
## Weight                         10.74694  1        3.278252
## factor(Man)                    14.73999  8        1.183126
## factor(Cylinders2)              2.74317  1        1.656252
## factor(Engine3):Horsepower     72.66560  7        1.358164
```

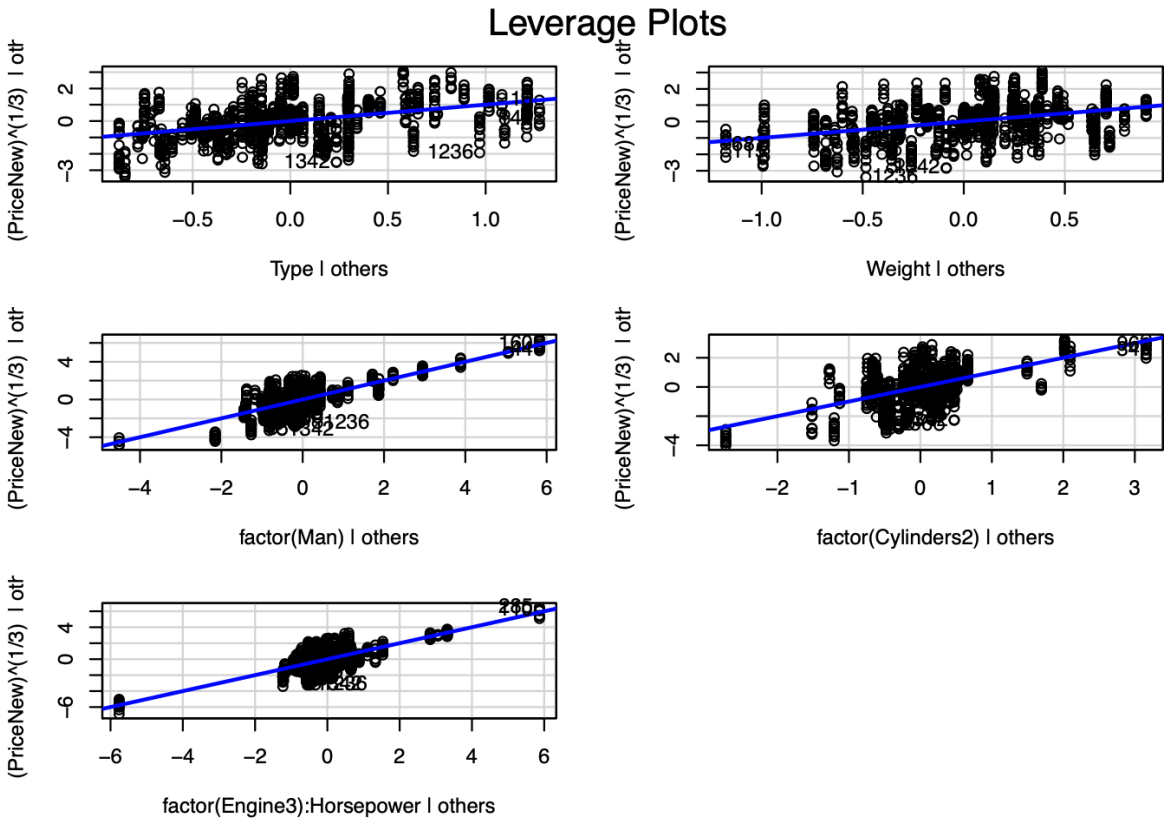
VIF Analysis of Final Model

As seen in the VIF analysis output of the final multiple linear regression model, we see that each predictor variable used in the model has a VIF score of less than 5, with all except one of the variables' scores being below 2. This tells us that there is no multicollinearity problem with the predictor variables being used in the model, and that the variables are independent and are not strongly correlated with each other. Thus, the assumption of independence is met, and the model seems to be valid in that sense.



Diagnostic Plots of Final Model

Looking at these diagnostic plots of the final multiple linear regression model used, it can also be seen that the assumptions of linearity, normality, and constancy of variance are being met with this model. The Residuals vs Fitted plot shows that the errors are consistently and randomly distributed above and below the red line, and the red line is almost completely horizontal, meaning that the errors are independent and uncorrelated with each other. The Normal Q-Q plot shows that the points very closely follow the diagonal line, meaning that the residuals are normally distributed. The Scale-Location plot similarly to the first plot shows an almost completely horizontal red line, meaning that the residuals and thus variance are constant as the fitted values increase. Finally, the Residuals vs Leverage plot shows that all the points are somewhat close to each other and the horizontal line, and there don't seem to be any particular points that are significantly further away from the rest, indicating that there are not any significantly bad leverage points. These four diagnostic plots corroborate the validity of our multiple linear regression model and show that the assumptions necessary to use a multiple linear regression model are adequately met.



Leverage Plots of Final Model

The leverage plots above show that all the variables being used in the multiple linear regression model are necessary and contribute significantly to the accuracy of the model. The non-zero slopes of the blue lines in each of the plots show that the variable is needed. Thus, we can conclude that all of the predictor variables in the model are needed and none of them are unnecessarily adding to the complexity of the model without an increase in prediction accuracy.

After using both the powerTransform and inverseResponsePlot functions on the model and experimenting with running t-tests on different versions of the model, the only transformation that seemed to significantly improve the accuracy of the model was taking the cubic root of the PriceNew variable. This suggested lambda was taken from the output of the

inverseResponsePlot function, which suggested a value close to 0.33. No other variable transformations were used in my final multiple linear regression model.

Using the EngineSize and Cylinders variables increased the complexity of the model a significant amount, and only some of the categories' β values were significant. So, to fix this, I created new versions of these variables, Cylinders2 and Engine3, to reduce the number of β's needed in the multiple linear regression model. Looking at the box plots for these variables showed that many of the factors of the variables had similar means, so these factors were combined so that the Cylinders2 variable only has 2 levels and the Engine3 variable only has 7, compared to the 7 and over 15 levels previously, respectively.

Additionally, after looking at the distribution of the residuals for this model's results, there were particular models and manufacturers that were consistently getting high residual values. So, to fix some of these large errors, I created a new variable called Man that identified cars of a particular manufacturer or make. This was extremely successful in decreasing the residuals of these instances and thus increasing the accuracy of the model as a whole. This new categorical variable ended up with a total of 9 levels.

Results

```
##
## Call:
## lm(formula = (PriceNew)^(1/3) ~ Type + Weight + factor(Engine3):Horsepower +
##     factor(Man) + factor(Cylinders2), data = train.Cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9058 -0.6519  0.0275  0.6352  2.7448
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.747e+01  5.404e-01  50.826  < 2e-16 ***
## TypeLarge                     2.001e-01  1.740e-01   1.150 0.250282
## TypeMidsize                   1.068e+00  1.278e-01   8.361  < 2e-16 ***
## TypeSmall                    -6.992e-01  1.375e-01  -5.087 4.33e-07 ***
## TypeSporty                    4.771e-01  1.280e-01   3.727 0.000204 ***
## TypeVan                      -6.295e-01  2.063e-01  -3.052 0.002334 **
## Weight                        2.346e-03  1.776e-04  13.205  < 2e-16 ***
## factor(Man)2                 -7.601e+00  3.557e-01 -21.368  < 2e-16 ***
## factor(Man)3                 -3.456e+00  3.574e-01  -9.672  < 2e-16 ***
## factor(Man)4                 -1.205e+01  5.751e-01 -20.953  < 2e-16 ***
## factor(Man)5                 -2.473e+00  4.720e-01  -5.239 1.96e-07 ***
## factor(Man)6                 -2.390e+00  3.534e-01  -6.765 2.24e-11 ***
## factor(Man)7                 -3.875e+00  5.370e-01  -7.217 1.03e-12 ***
## factor(Man)8                 -3.632e+00  3.565e-01 -10.189  < 2e-16 ***
## factor(Man)9                 -6.559e+00  2.652e-01 -24.731  < 2e-16 ***
## factor(Cylinders2)2          -4.382e+00  1.932e-01 -22.679  < 2e-16 ***
## factor(Engine3)1:Horsepower   1.505e-02  2.252e-03   6.680 3.90e-11 ***
## factor(Engine3)2:Horsepower   5.761e-02  3.013e-03  19.120  < 2e-16 ***
## factor(Engine3)3:Horsepower   3.021e-02  2.734e-03  11.051  < 2e-16 ***
## factor(Engine3)4:Horsepower   3.468e-02  1.748e-03  19.840  < 2e-16 ***
## factor(Engine3)5:Horsepower   1.774e-02  2.313e-03   7.672 3.94e-14 ***
## factor(Engine3)6:Horsepower -1.690e-02  3.119e-03  -5.420 7.41e-08 ***
## factor(Engine3)7:Horsepower   1.655e-02  1.502e-03  11.013  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.053 on 1027 degrees of freedom
## Multiple R-squared:  0.9305, Adjusted R-squared:  0.929
## F-statistic: 624.6 on 22 and 1027 DF,  p-value: < 2.2e-16
```

Summary of Final MLR

The summary of the multiple linear regression model above shows its 23 β coefficients, all of which but one have an extremely small p-value and are therefore very significant. A total of three categorical variables, one numerical variable, and one interaction term between a categorical and numerical variable were used. We also note from the summary output that the R-Squared value of the model is 0.9305 and the Adjusted R-Squared is 0.929.

```
## Analysis of Variance Table
##
## Response: (PriceNew)^(1/3)
##                              Df Sum Sq Mean Sq F value    Pr(>F)
## Type                          5 8716.5 1743.30 1572.27 < 2.2e-16 ***
## Weight                        1 2662.6 2662.60 2401.37 < 2.2e-16 ***
## factor(Man)                   8 1593.0  199.13  179.59 < 2.2e-16 ***
## factor(Cylinders2)            1 1061.4 1061.36  957.23 < 2.2e-16 ***
## factor(Engine3):Horsepower    7 1202.3  171.76  154.90 < 2.2e-16 ***
## Residuals                  1027 1138.7    1.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA Table of Final MLR

The ANOVA table above shows the five different predictor variables used in the final model, and how much each variable contributes to explaining the car's price. We see from the table that each variable has a very small p-value, so each variable has a significant contribution to the Residual Sum of Squares and to increasing the R-Squared value of the model as a whole. We also note that the Type variable has the largest contribution to the Sum Sq, followed by Weight, and finally Man, Cylinders2, and Engine3:Horsepower, which all have similar levels of contribution. Most importantly, this tells us again that we have no unnecessary predictor variables in our multiple linear regression model.

Discussion

The summary of our model shows that the residuals are centered around 0 and that the first quartile and third quartile, as well as the min and max are roughly equidistant from the median. This is a good indication of our model having randomly and independently distributed residuals.

Analyzing the coefficients of our model as seen in its summary, we can draw the following conclusions:

- The baseline for a car's price is $27.47^3 = 20{,}728.89$.
- A car having Type Large, Midsize, or Sporty increases the price by the respective coefficients listed in the summary, while a car having Type Small or Van decreases its price by about the same amount. A car having Type Sporty on average sees the largest price increase compared to other types, followed by Large and then Midsize.
- A car's price increases as its weight increases.
- The various 'Man' categories subtract a different amount from a car's predicted price, so the 1 level of the Man variable must be the group of cars that have the highest price.
- A car with 8 or rotary cylinders sees a price increase of $4.382^3 = 84.14$ on average compared to a car with 2, 4, or 6 cylinders.
- An increase in horsepower results in a variety of increases or decreases in price based on the car's EngineSize. More analysis on this could be done by someone with more background knowledge on car engines to gain further insights.

As mentioned previously, the ANOVA table shows that Type by far has the largest contribution in explaining variation of car prices. Combining this with the results found above

shows us that using these mean prices for each type of car is a very good prediction baseline when predicting a car's price.

Weight also has a large contribution to the Residual Sum of Squares and is also a strong predictor variable as it is a numerical variable and thus only results in one β coefficient. Thus, this variable significantly contributes to the accuracy of the model without increasing its complexity too much.

The Man, Cylinders2, and Engine3:Horsepower variables still have a significant contribution to the model's accuracy, though smaller than the previous two mentioned.

The insights obtained from the summary and ANOVA table of our multiple linear regression model prove extremely useful in developing a good understanding of the factors that affect the pricing of a car and give Geely Auto a strong foundation to begin planning their manufacturing unit and business plan.

Limitations and Conclusions

As just mentioned in the previous section, the Man, Cylinders2, and Engine3:Horsepower variables do not have as large of a contribution to the model's accuracy as the Type and Weight variables. As the Cylinders2 variable only has two levels, it does not add to the complexity of the model too much, but the Man and Engine3:Horsepower variables have nine and eight levels respectively. So, one limitation of my final model is that there may be better predictor variables to use that would not have raised the complexity of the model so much. Alternatively, there may have been more effective ways to create the Engine3 and Man variables, such as combining some of the levels that have more similar mean prices. The number of β coefficients added to the model through these two variables may or may not be the ideal decision in the long run, as there may be more effective variables to use or create in their place.

Similarly, the method of creating the Man variable could likely benefit from some improvements as well. I created the variable by identifying the models and manufacturers of the observations with the largest residuals, but could have added other models and manufacturers with similar mean prices to the various levels. While this would not have reduced the number of β coefficients in the model, it would have likely upped its accuracy a significant amount. Additionally, it could be worth exploring whether using a car's model, make, or manufacturer is most effective, as I ended up using a combination of both model and manufacturer in my Man variable. It would also be beneficial to explore if there are other unused variables that account for these differences in these edge cases, and if including these variables may be a better strategy than creating the Man variable.

One final improvement that could be made to the model, though somewhat small, would be to create a new Type variable that does not include the 'Large' level. This is because in the

summary of the model, the TypeLarge coefficient is not significant. This would easily reduce the number of β coefficients in the model by one without much drop in accuracy and result in a slight improvement in complexity.

In conclusion, Type seems to be an extremely effective predictor of a car's price, especially when combined with Weight. Using a car's manufacturer, model, or make to create a new variable out of more non-ordinarily priced makes is a good way of accounting for these more 'outlier' values and allowing the model to predict these instances more accurately. The interaction term between a car's EngineSize and Horsepower, as well as its Cylinders are also good predictors of its price; however, for these last three variables mentioned, there likely may be alternative methods that could do the same job.

I would advise Geely Auto to explore the pricing of cars in the American market differently based on a car's type and do different analyses for vans, sporty cars, large cars, midsize cars, and small cars. It may also be beneficial to create different multiple linear regression models within each of these subgroups to see if different variables and factors are more or less significant predictors of price with these different types of cars.

References

Almohalwas, Akram. Statistics 101A Chapter 6.

Sheather, Simon J. A Modern Approach to Regression with R. Springer Texts in Statistics, 2009.