**DSC 423: Data Analysis and Regression**

**Assignment 04: Model Building**

**Name:** Adarsh Shankar

**Student ID:** 2117611

**Honour Statement:** “I have completed this work independently. The solutions given are entirely my own work." Your submission must be submitted as a PDF”.

**Q1.** CARPRICE. Find on the D2L a car price dataset. Use R to perform a regression analysis on the dataset Your submission should take the form of a technical report and should consider the following:

1. (10 pts.) Paste your final model into your submission (just the R output).

Graphical user interface, text, application

Description automatically generated

1. (10 pts.) Describe the model building process through which you generated this model.

* I developed a number of models and included a large number of second order and interaction factors to gauge the variable's significance. For every model I tried, I did, however, get an adjusted R-squared value of 1.

A variation of R-squared that takes into consideration variables in a regression model that are not significant is called adjusted R-squared. When compared to a model with more input variables, the adjusted R-squared is lower, indicating that the additional variables do not enhance the model.

In comparison to a model with more input variables, a higher adjusted R-squared indicates that the additional input variables are improving the model. I too found an almost perfect correlation between price and square meters using the cor(CARPRICE) at first.

This made it a perfect model in this instance. I attempted to drop square meters and construct a model in an effort to learn. At the conclusion of the task, I've attached the work.

1. (10 pts.) What significant second-order terms did you find, if any? Did you try all second-order terms? Did you look at scatter plots to determine which second-order terms to evaluate? Discuss the benefits and drawbacks of these two strategies.

* The correlation revealed a strong positive correlation between pricing and car width, curb weight, and horsepower. When these 3 terms were included in the second order model, the adj R2 jumped dramatically from 81.55% to 82.82%, and the p values also dropped significantly. Other phrases did not much improve the situation, and some even had detrimental impacts.

The optimal second order terms were discovered to be car width, curb weight, and horsepower.

Horsepower, curb weight, and car width had the strongest positive relationships in terms of the corelation. The relationships between other terms weren't very strong.

1. (10 pts.) What significant interaction terms did you find, if any? Did you try all combinations of interaction terms? Do you think that is an appropriate strategy? What happens to the number of interaction terms as the number of independent terms increases?

* "HorsePowerxcompressionratio" and "HorsePowerxstroke" in the model increased adjusted R-squared from 82.82% to 83.53%, improved F-test result, and reduced pvalues for coefficients, indicating enhanced model performance and increased significance of the interaction terms.

The model did not show improvement with other attempted interaction terms such as "carwidth-curbweight," "carwidth-horsepower," and "curbweight-carwidth." The adjusted R-squared, F-test result, and p-values did not show significant changes, suggesting that these interaction terms may not have a strong impact on the model's performance.

In a technical sense, horsepower and stroke are interconnected concepts in the context of the model. Considering the interaction between these two variables may have been meaningful as it could capture the combined effect of both variables on the model's performance, leading to improved results in terms of adjusted R-squared, F-test, and pvalues.

1. (10 pts.) Discuss your final model. Evaluate the t-tests, F-Test and adj-R2 accordingly. Do you think this is a “good” model? Explain.

* The second order terms and interaction terms' t tests pass muster, hence this model is valid.

The regression model's independent variables are responsible for 83.53% of the price fluctuation. The null hypothesis can be rejected and the alternative accepted since at least one beta is not equal to 0.

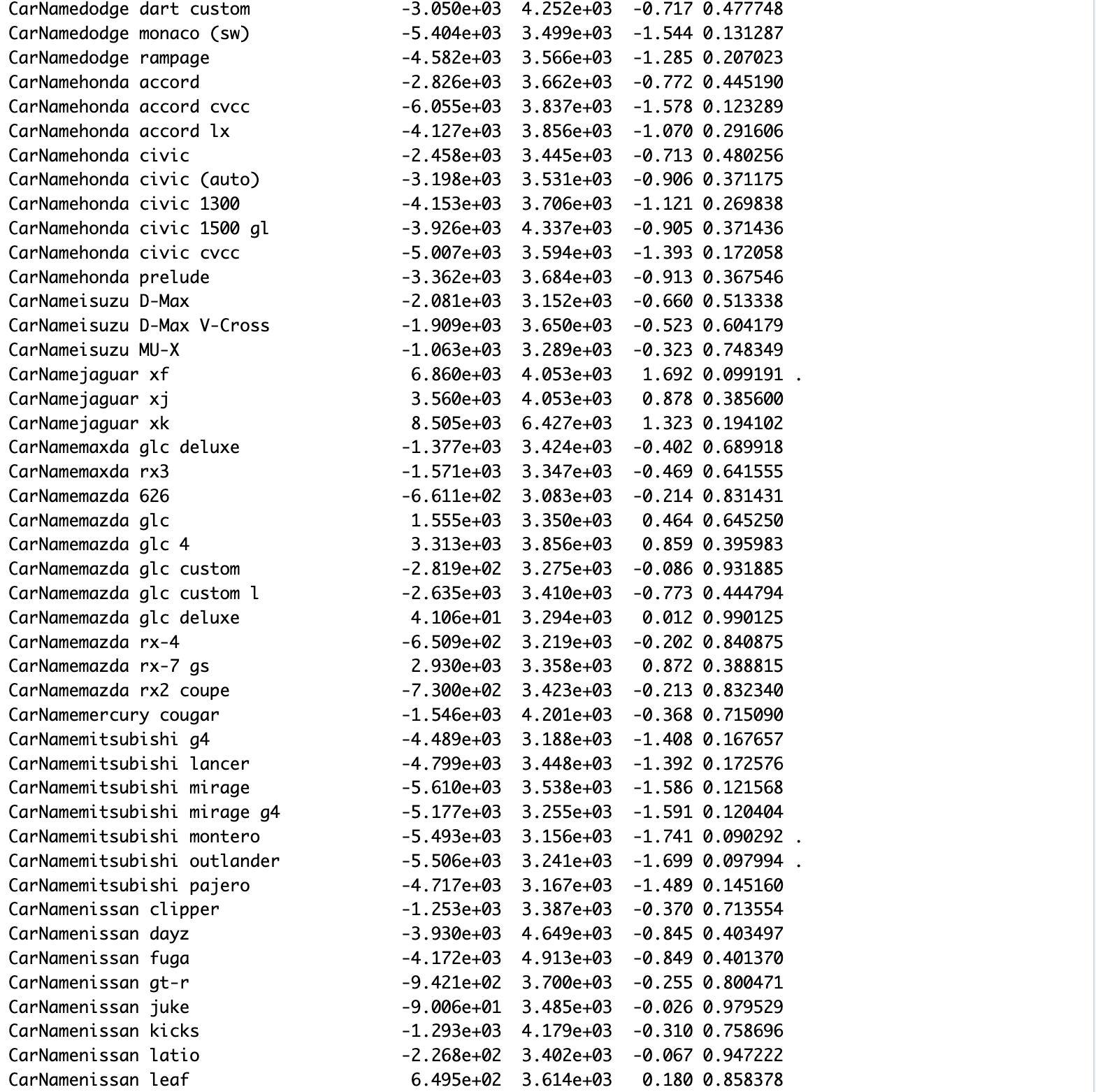
By that we can say that

* 1. 104.4 on 10 and 194 DF for the F-test
  2. Adj R2: 83.53

1. Include your code an appendix.

Table

Description automatically generated



Table

Description automatically generated

Table

Description automatically generated