Sean McLean

ALY 6020

Module 2

Mid-Week Project

Introduction

The focus of the module assignment is to find the best independent variables that will help best predict the price of cars. An Excel dataset is provided with 26 variables and 205 cells with the price variable being the dependent variable that will be examined. Initial steps include downloading the necessary libraries and packages in Python, importing the dataset, analyzing the summary statistics of the dataset, and data cleaning for regression analysis. Some of the categorical attributes were converted to numerical values so the OLS model can be processed for proper analysis of each variable. The three most relevant and effective variables are selected from the dataset to answer the question that pertains to the best prediction of car prices.

1. **What were the three most significant variables?:** After executing the codes, the OLS regression results showed that the top variables that will predict the price of the cars are engine location, engine size, and car width. These were based off two factors, the p-value, the correlation coefficient, and the coefficient of determination, and therefore were not the top three variables based just off the lowest p-values. The variables engine location and engine size were two of the four variables that had p-values of 0.00 and had positive correlation coefficients. The other two variables with p-values of 0.00 both contained negative correlation coefficients which would have had negative impacts on the price variable. The car width variable had a p-value of 0.020 and was the next variable in line of lowest p-values among the variables that had a strong positive correlation coefficient.

2. **Of those three, which had the greatest positive influence on car prices?:** From the metrics, the engine location variable looked to be the most significant variable that had the lowest p-value (0.00) and highest positive correlation coefficient at 11,090. When it was plotted however, it did not show a strong linear relationship which could be because it had a lot of outliers that could affect the regression analysis (See Appendix A). The variable that seems to have the greatest positive influence on car prices is engine size, showing a strong positive linear relationship with only a few outliers contained in the scatterplot (See Appendix B).

   The second-best variable would probably be the car width variable which shows a positive linear relationship in relation to the dependent variable but has a lot more outliers than the engine size variable (See Appendix C). The fourth best variable by statistical standards was car height which was plotted to see the relationship between the two variables. Looking at the plot there is only a minimal positive relationship which is not surprising considering the car height p-value was barely under 0.05 and did not have a large correlation coefficient (See Appendix D).

3. **How accurate was the model?:** The OLS regression results that were executed had a coefficient of determination value of 0.90 which indicates that the model is very accurate and contains very few errors. This value can change depending on which variables are removed from the model that will provide the best results for regression analysis.

Conclusion

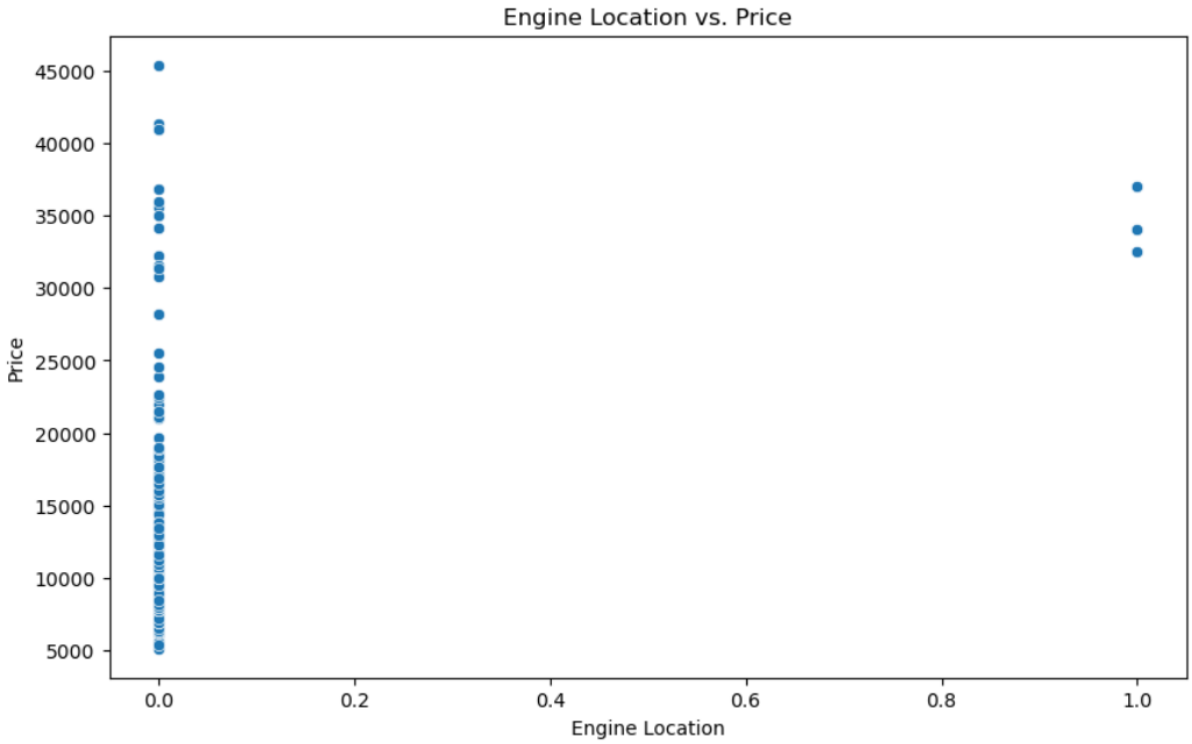
The three variables that had the strongest positive relationship to predicting the prices of cars pertained to engine characteristics and the size of the car overall. This would suggest that engines in different capacities have a major impact on how much a vehicle will cost. Car width and height in relation to price could indicate that car size will affect the vehicle's price since there is more material involved in constructing larger vehicles. The model was accurate overall in showing that these variables have the strongest relationship with the price variable.

References

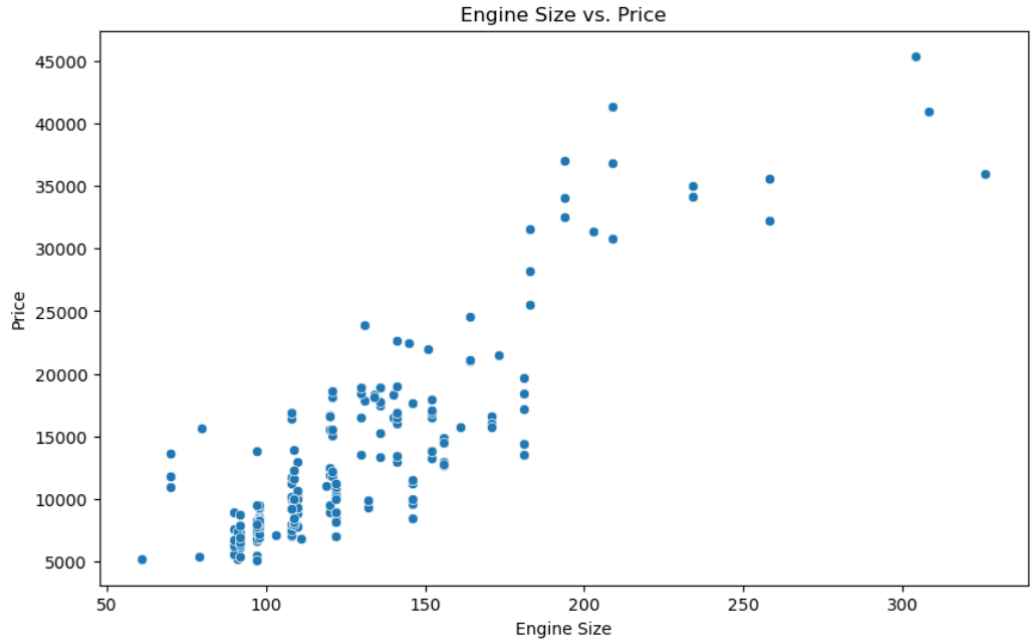
Anaconda Distribution. (2024). Anaconda (Version 4.12.3). Retrieved July 21, 2024, from https://www.anaconda.com

Appendix

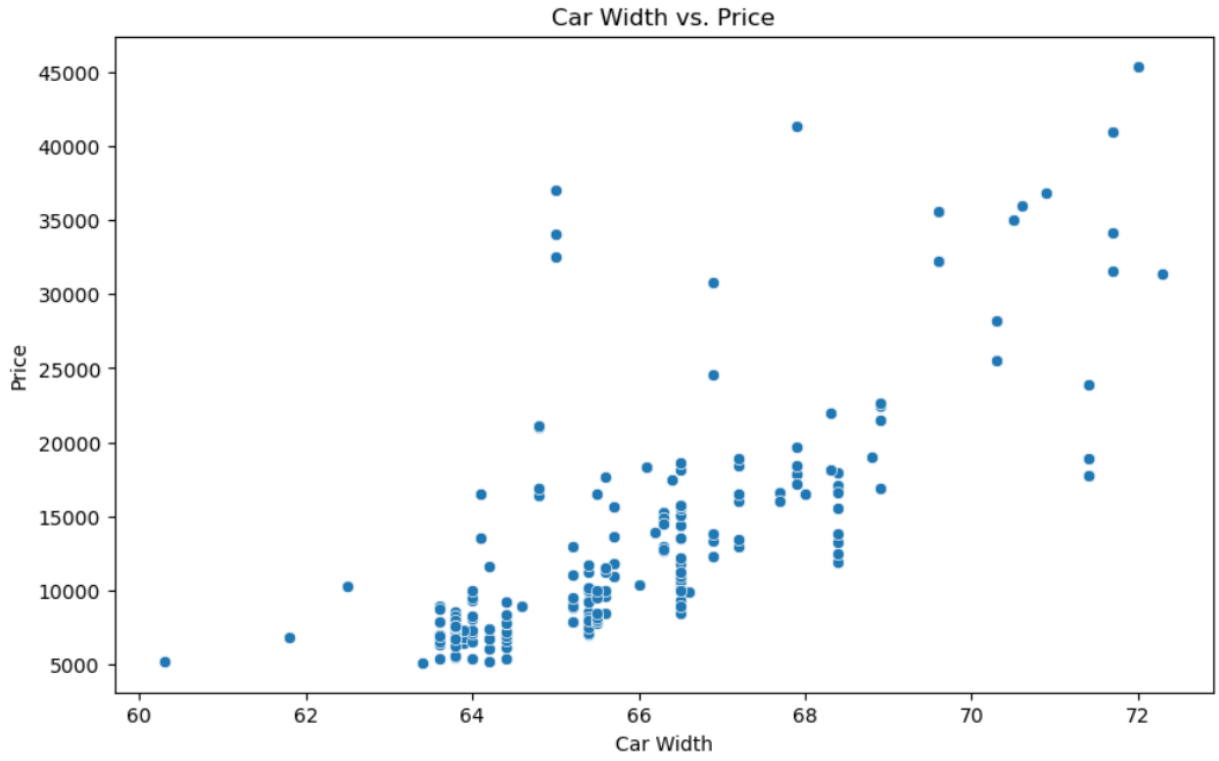
Appendix A



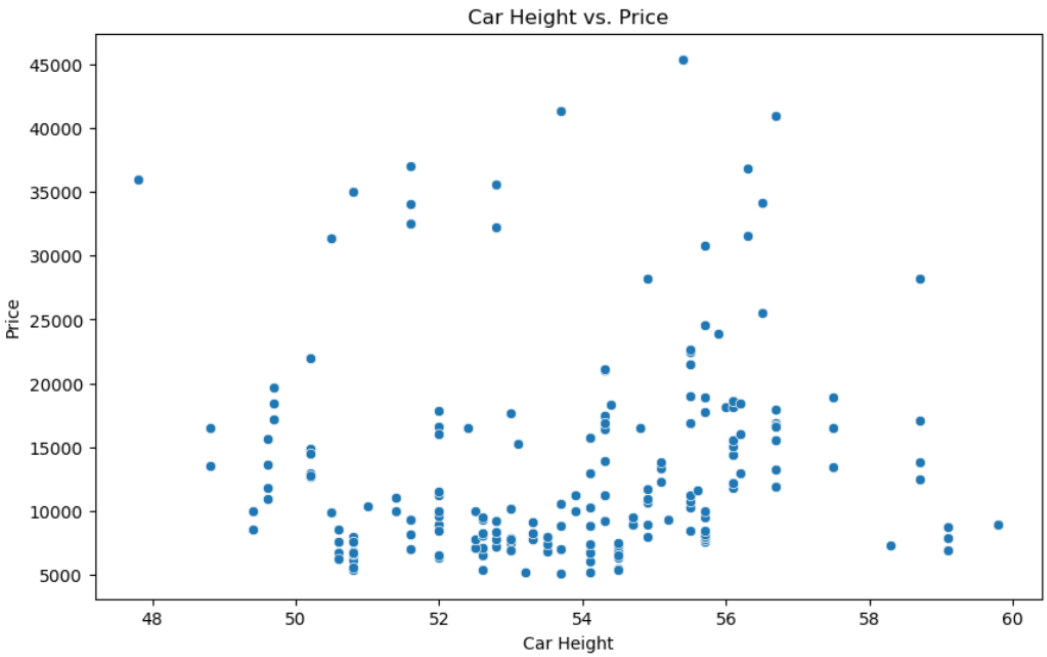
Appendix B



Appendix C



Appendix D


Car Height vs. Price
Price
Car Height
45000
40000
35000
30000
25000
20000
15000
10000
5000
48
50
52
54
56
58
60