An Analysis of Variables Affecting Car Prices in the US

Jun Zhao

Case Western Reverse University

Cleveland Ohio

jxz1395@case.edu

ABSTRACT

This paper presents an analysis of the variables affecting car prices in the US. The study uses descriptive statistics such as correlation, covariance, and mean to identify the most important factors influencing car prices. The study also employs regression analysis using REF, VIF, and OLS methods to determine the impact of different variables on car prices. The results indicate that the most important variables that influence car prices are horsepower, car width, hatchback (representing a kind of car body), and high-end (representing luxury car brands, such as Porsche, BMW, etc.). The study provides several plots to illustrate the relationships between the different variables and car prices. These findings can be useful for car manufacturers, dealers, and buyers in making informed decisions about pricing and purchasing cars.

Keywords: car prices, variables, OLS, VIF, REF, statistical analysis

1 Introduction and Background

The US automobile market is one of the largest automobile markets in the world, and consumers are highly concerned about the price of cars, which involves multiple factors. So, this project aims on determining which factors determine the price of cars in the United States of America. The purpose of this study is to provide insights and understanding into the factors that influence the price of cars in the United States, which can be useful for car manufacturers, dealers, and consumers. By identifying the key factors that determine the price of cars, this study aims to help stakeholders make informed decisions about car pricing, marketing, and purchasing.

This study falls under the field of economics and is a quantitative research design. The goals of the study can be answered using statistical analysis, which allows us to identify significant relationships between price and various factors. To achieve this goal, we will analyze a dataset of prices and corresponding variables such as 'housepower' and 'fueltype' indicators. We will use statistical analysis techniques to identify correlations and patterns in the data and develop models to predict price based on the independent variables.

2 The Sample Study

The dataset is downloaded from Kaggle. This database contains 204 data entries, including the following independent variables: **'CarName', 'fueltype', 'aspiration', 'doornumber', 'carbody', 'drivewheel', 'enginelocation','wheelbase', 'carlength', 'carwidth', 'carheight', 'curbweight', 'enginetype','cylindernumber', 'enginesize', 'fuelsystem', 'boreratio', 'stroke','compressionratio', 'horsepower', 'peakrpm', 'citympg', 'highwaympg', and thedependent variable 'price'.**

The independent variable **'CarName'** contains data like: 'bmw x1', 'audi 5000s (diesel)', and 'dodge monaco (sw)' and so on. From this, we could get which company produces this car. The independent variable **'fueltype'** refers to the type of fuel that a car uses to power its engine and contains data like 'gas', 'diesel'. The independent variable **'aspiration'** refers to the process of drawing in air into the engine's combustion chamber for mixing with fuel and ignition to produce power and contains data like 'std(represents standard)', 'turbo(represents turbocharged engine)'. The independent variable 'doornumber' is the number of the door on the car and it contains data like 'two', 'four'. The independent variable **'carbody'** refers to the outer shell or structure of a car that encloses the passenger compartment and other mechanical components and it contains data like 'sedan'', 'hatchback', 'wagon', and so on. The independent variable **'drivewheel'** refers to the wheels in a vehicle that are responsible for transferring power from

the engine to the ground to move the vehicle forward or backward and it contains '4wd', 'rwd', 'fwd'. The independent variable **'enginelocation'** refers where the engine is situated in the vehicle, and it contains data 'front' and 'rear'. The independent variable **'wheelbase'** is the horizontal distance between the centers of the front and rear wheels. The independent variable **'carlength'** is the length of the car. The independent variable **'carwidth'** is the width of the car. The independent variable **'carheight'** is the height of the car. The independent variable **'carweight'** is the weight of the car. The independent variable **'enginetype'** contains data like 'ohc', 'dohc', 'l' and so on. The independent variable '**cylindernumber'** is the number of cylinders of the car. The independent variable **'enginesize'** is the size of the engine. The independent variable **'fuelsystem'** contains data like 'mpfi(stands for Multi-Point Fuel Injection)', '2bbl(represents two-barrel carburetor)','1bbl(stands for one-barrel carburetor)' and so on. The independent variable **'boreratio'** is the bore-stroke ratio of the car. The independent variable **'stroke'** is the distance that the piston travels up and down inside the engine cylinder. The independent variable **'compressionratio'** is the ratio of the volume of the combustion chamber (when the piston is at the bottom of its stroke) to the volume of the same chamber when the piston is at the top of its stroke. The independent variable **'horsepower'** refers to the power output of the engine and is a measure of how much power the engine can produce to move the car forward. The independent variable **'peakrpm'** refers to Peak RPM (Revolutions Per Minute) is the maximum rotational speed of a car's engine crankshaft, measured in revolutions per minute. The independent variable 'The independent variable **'citympg'** refers to the estimated city fuel economy of a car, which is the number of miles per gallon (mpg) a car can travel in city driving conditions. The independent variable **'highwaympg'** refers to the estimated city fuel economy of a car, which is the number of miles per gallon (mpg) a car can travel in highway driving conditions. The dependent variable 'price' is the price of this car.

3 Statistical Analysis

3.1 Data preprocessing

There are some independent variables in this dataset, that are not well classified for use in the study, so I have treated them. For example, I find that the independent variable **'doornumber'** only have two kinds of value: 'two' or 'four' which could be converted to number, so I changed this column to number using python. I also combine variables **'citympg'** and **'highwaympg'** to **'fueleconomy'**. In this project, **'fueleconomy'**= 0.55 * **'citympg'**+0.45 * **'highwaympg'**.

When we talk about the factors that determine the price of a car, we first consider the brand of a car. There is no doubt that high-end car brands are indeed higher in terms of pricing. From the datasets I found some companies with names written in two ways, such as 'maxda' and 'mazda', 'porcshce' and 'porsche, and so on. So before starting the experiment, I standardized these names to the ones we use today while ensuring the authenticity of the data. What's more, I also plot the box figure to better understand how the brand influence the price of a car in US. The box plot results indicate that the brand of a car can have an impact on its price. To further analyze this relationship, the independent variable 'CarName' was divided into three categories: 'Budget', 'Medium', and 'Highend', based on the average price of each brand. This categorization allows for a clearer comparison of the different brands and their corresponding price ranges.

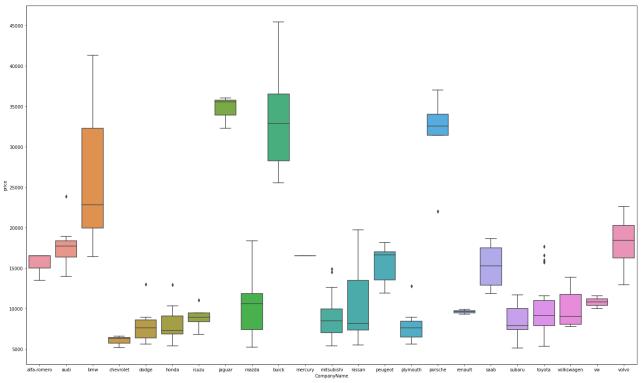

Figure 1: Box plot for varaiable 'CarName'

In addition, there are a significant number of variables in the database whose data types consist only of strings, which means it is not possible to calculate their mean values and the variances, and the correlation and the covariance between them and the dependent variable price. They are '**CarName', 'fueltype', 'aspiration', 'carbody', 'drivewheel', 'cylindernumber'** and **'enginetype'.** So

before training the model, I add dummy variables for these variables. For example, for the 'carbody' variable I divided it into 'sedan', 'wagon', and 'hatchback'. Dummy variables are added to a regression model to account for the effects of these categorical variables. This is necessary because regression models require numerical input, and categorical variables cannot be directly input as numerical data. By creating dummy variables, we can assign numerical values to each category of the categorical variable, making it possible to include the variable in the regression analysis.

3.2 Descriptive statistics

The mean of the dependent variable **price** is 13276.710571. And the mean of these independent variables (**symbolling, doornumber, wheelbase, carlength, carwidth, carwidth, curbweight, enginesize, boreratio, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg.**) is 0.834146, 3.121951, 98.756585, 174.049268, 65.907805, 53.724878, 2555.565854, 126.907317, 3.329756, 3.255415, 10.142537, 104.117073, 5125.121951, 25.219512, 30.751220 respectively.

The correlation between dependent variable **price** and independent variables (**symbolling, doornumber, wheelbase, carlength, carwidth, carwidth, curbweight, enginesize, boreratio, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg.**) is -0.079978, 0.031835, 0.577816, 0.682920, 0.759325, 0.119336, 0.835305, 0.874145, 0.553173, 0.079443, 0.067984, 0.808139, -0.085267, -0.685751, -0.697599 respectively. The heatmap according to the correlation is as following:

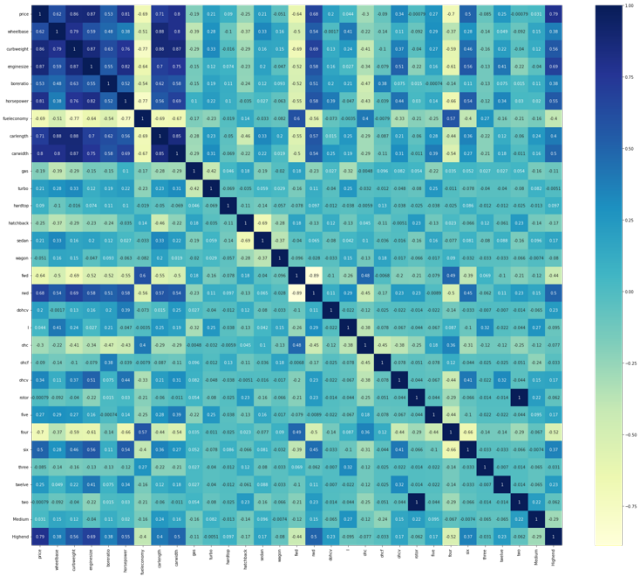

Figure 2: Heatmap of correlation between each variable

The covariance between dependent variable **price** and independent variables (**symbolling, doornumber, wheelbase, carlength, carwidth, carwidth, curbweight, enginesize, boreratio, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg.**) is -795.669155, 253.046906, 27797.019319, 67309.126836, 13013.101947, 2329.554864, 3.474565e+06, 290808.157690, 1196.917731, 199.027188, 2157.255569, 255301.163227, -324916.262938, -35840.247445, -38378.258647 respectively.

The variance of the dependent variable **price** is 6.382176e+07. And the mean of these independent variables (**symbolling, doornumber, wheelbase, carlength, carwidth, carwidth, curbweight, enginesize, boreratio, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg.**) is 1.550789e+00, 9.899570e-01, 3.626178e+01, 1.522087e+02, 4.601900e+00, 5.970800e+00, 2.711079e+05, 1.734114e+03, 7.335631e-02, 9.834309e-02, 1.577710e+01, 1.563741e+03, 2.275153e+05, 4.279962e+01, 4.742310e+01 respectively.

In addition to performing calculations, I also created visual representations of the relationship between the different variables and the dependent variable (price) using box plots, scatter plots, and pairplot plots. These plots help to provide a clearer understanding of how the variables are related to the price.

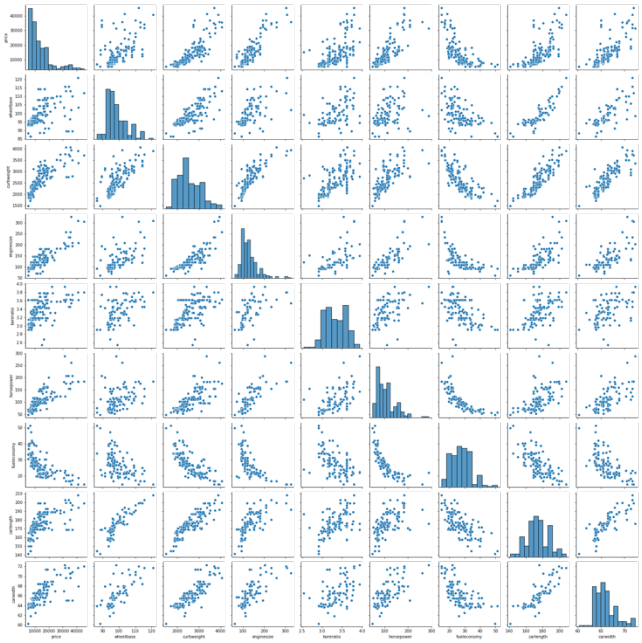

Figure 3: Pair plot of each variable

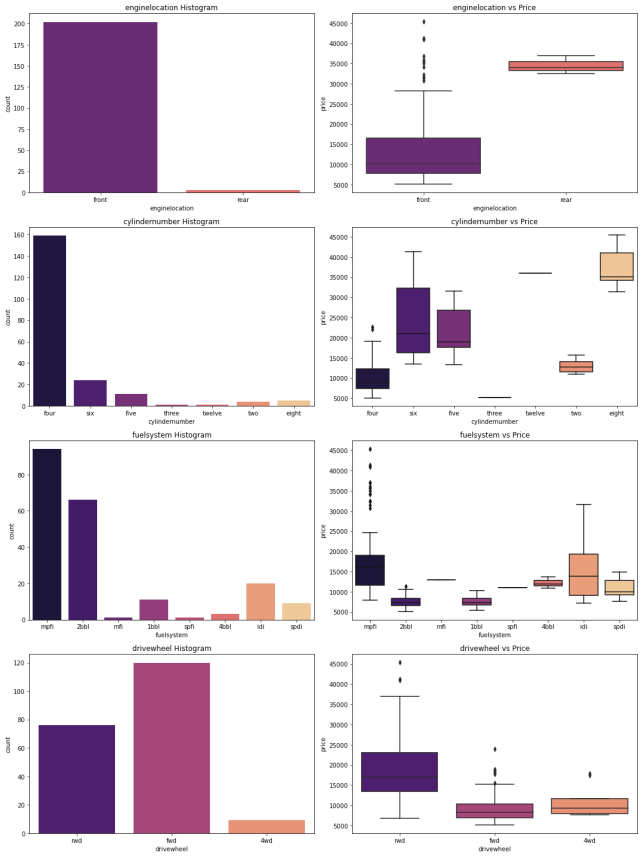

Figure 4: Box plots and histogram of some variables

After analyzing a series of graphs and calculating correlation, covariance, and mean for several variables, I determined that **'fueltype', 'aspiration', 'carbody', 'drivewheel', 'wheelbase', 'curbweight', 'enginetype', 'cylindernumber', 'enginesize', 'boreratio', 'horsepower', 'fueleconomy', 'carlength', 'carwidth', and 'carsrange'** have a significant impact on car prices. Thus, I decided to use these variables as inputs for the regression model training to obtain a reliable model.

3.3 Regression analysis

After adding dummy variables, I applied the Recursive Feature Elimination (RFE[1]) method to the dataset and identified the top 10 most significant variables, which are **'curbweight', 'horsepower', 'fueleconomy', 'carwidth', 'hatchback', 'sedan', 'wagon', 'dohcv', 'twelve', and 'Highend'.** RFE is a feature selection technique used in machine learning to identify the most important features in a dataset. It works by recursively removing features from the dataset and building a model on the remaining features, and then ranking the importance of the removed features based on how much they contribute to the model's performance. RFE can be useful in reducing the dimensionality of a dataset and improving the performance of machine learning models by reducing overfitting and noise in the data. However, it is important to choose the right number of features to select and to carefully evaluate the performance of the model using the selected features to ensure that it is still accurate and reliable.

OLS[2], or Ordinary Least Squares, is a method for estimating the parameters of a linear regression model. It is a widely used technique for modeling the relationship between a dependent variable and one or more independent variables in a dataset. In OLS, the goal is to find the line that best fits the data points by minimizing the sum of the squared differences between the predicted and actual values of the dependent variable. The equation for a simple linear regression model with one independent variable can be written as: y = b0 + b1 * x + e. where y is the dependent variable, x is the independent variable, b0 and b1 are the intercept and slope coefficients, respectively, and e is the error term or residual. The OLS method estimates the values of b0 and b1 that minimize the sum of the squared residuals for the given data. This is done by calculating the partial derivatives of the sum of squared

residuals with respect to the intercept and slope coefficients, and then solving for the values that set these derivatives equal to zero.

However, when the independent variables in a linear regression model are correlated, the OLS method can produce biased and unreliable estimates of the regression coefficients. To address this issue, the Variance Inflation Factor (VIF[3]) technique is used to measure the extent of multicollinearity among the independent variables. VIF, or Variance Inflation Factor, is a measure of how much the variance of the estimated regression coefficients are inflated due to collinearity in the independent variables of a multiple linear regression model. The VIF for each independent variable is calculated as the ratio of the variance of the estimated regression coefficient when that variable is included in the model, to the variance of the estimated regression coefficient when that variable is excluded from the model. A VIF value of 1 indicates that there is no collinearity, while values greater than 1 suggest the presence of collinearity. A commonly used threshold for identifying high collinearity is a VIF value of 5 or greater.

Secondly, I employed a combination of the Ordinary Least Squares (OLS) method and the Variance Inflation Factor (VIF) technique to train the linear regression model. I conducted a series of iterations in which I identified the variables that had little or no impact on the price of the car based on their P-values and VIF values in the OLS method, and then eliminated them from the model. By repeating this process, I continually optimized the model until I generated a final model that only included the most significant variables that had a significant impact on the price of the car.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.929
Model:	OLS	Adj. R-squared:	0.923
Method:	Least Squares	F-statistic:	172.1
Date:	Sun, 23 Apr 2023	Prob (F-statistic):	1.29e-70
Time:	13:22:35	Log-Likelihood:	205.85
No. Observations:	143	AIC:	-389.7
Df Residuals:	132	BIC:	-357.1
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0947	0.042	-2.243	0.027	-0.178	-0.011
curbweight	0.2657	0.069	3.870	0.000	0.130	0.402
horsepower	0.4499	0.074	6.099	0.000	0.304	0.596
fueleconomy	0.0933	0.052	1.792	0.075	-0.010	0.196
carwidth	0.2609	0.062	4.216	0.000	0.138	0.383
hatchback	-0.0929	0.025	-3.707	0.000	-0.143	-0.043
sedan	-0.0704	0.025	-2.833	0.005	-0.120	-0.021
wagon	-0.0997	0.028	-3.565	0.001	-0.155	-0.044
dohcv	-0.2676	0.079	-3.391	0.001	-0.424	-0.112
twelve	-0.1192	0.067	-1.769	0.079	-0.253	0.014
Highend	0.2586	0.020	12.929	0.000	0.219	0.298

Omnibus:	43.093	Durbin-Watson:	1.867
Prob(Omnibus):	0.000	Jarque-Bera (JB):	130.648
Skew:	1.128	Prob(JB):	4.27e-29
Kurtosis:	7.103	Cond. No.	32.0

Figure 5: Result 1 of OLS Regression.

From the Figure, we can find that variables **'hatchback', 'sedan', 'wagon', 'dohcv', 'twelve'**(dummy variable of **cylindernumber**) have a negative impact on the price of a car. And all the other variables have a positive impact on the price of a car. Notice the column P>|t|, the value of variable 'twelve' is largest, so drop this variable for the second experiment. Notice: the **'const'** term represents the intercept or constant term in the linear regression equation.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.927
Model:	OLS	Adj. R-squared:	0.922
Method:	Least Squares	F-statistic:	187.9
Date:	Sun, 23 Apr 2023	Prob (F-statistic):	4.25e-71
Time:	13:22:39	Log-Likelihood:	204.17
No. Observations:	143	AIC:	-388.3
Df Residuals:	133	BIC:	-358.7
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0764	0.041	-1.851	0.066	-0.158	0.005
curbweight	0.2756	0.069	3.995	0.000	0.139	0.412
horsepower	0.3997	0.069	5.824	0.000	0.264	0.535
fueleconomy	0.0736	0.051	1.435	0.154	-0.028	0.175
carwidth	0.2580	0.062	4.137	0.000	0.135	0.381
hatchback	-0.0951	0.025	-3.766	0.000	-0.145	-0.045
sedan	-0.0744	0.025	-2.983	0.003	-0.124	-0.025
wagon	-0.1050	0.028	-3.744	0.000	-0.160	-0.050
dohcv	-0.2319	0.077	-3.015	0.003	-0.384	-0.080
Highend	0.2565	0.020	12.743	0.000	0.217	0.296

Omnibus:	48.027	Durbin-Watson:	1.880
Prob(Omnibus):	0.000	Jarque-Bera (JB):	159.802
Skew:	1.231	Prob(JB):	1.99e-35
Kurtosis:	7.556	Cond. No.	29.6

Figure 6: Result 2 of OLS Regression

Again, because the P>|t|-value of the variable 'fueleconmy' is too large, we decide to discard it for the next experiment.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0305	0.026	-1.165	0.246	-0.082	0.021
curbweight	0.2593	0.068	3.796	0.000	0.124	0.394
horsepower	0.3469	0.058	5.964	0.000	0.232	0.462
carwidth	0.2488	0.062	3.995	0.000	0.126	0.372
hatchback	-0.0922	0.025	-3.650	0.000	-0.142	-0.042
sedan	-0.0711	0.025	-2.850	0.005	-0.120	-0.022
wagon	-0.1047	0.028	-3.721	0.000	-0.160	-0.049
dohcv	-0.1968	0.073	-2.689	0.008	-0.342	-0.052
Highend	0.2610	0.020	13.083	0.000	0.222	0.301

Omnibus:	48.637	Durbin-Watson:	1.909
Prob(Omnibus):	0.000	Jarque-Bera (JB):	161.444
Skew:	1.250	Prob(JB):	8.77e-36
Kurtosis:	7.566	Cond. No.	27.2

Figure 7: Result 3 of OLS Regression

From the result, we can find that no variable has an excessive P>|t| value, and this is when VIF comes into play. Check the VIF value we can find that the value of 'curbweight' is quite large, so we will drop this value.

	Features	VIF
0	const	26.90
1	curbweight	8.10
5	sedan	6.07
4	hatchback	5.63
3	carwidth	5.14
2	horsepower	3.61
6	wagon	3.58
8	Highend	1.63
7	dohcv	1.46

Figure 8: The result3 of VIF

Repeating the above steps until the P>|t| values of all variables and the VIF values are within the acceptable range, we obtain our final model, here is the result of final model:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.899
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	308.0
Date:	Sun, 23 Apr 2023	Prob (F-statistic):	1.04e-67
Time:	13:23:22	Log-Likelihood:	181.06
No. Observations:	143	AIC:	-352.1
Df Residuals:	138	BIC:	-337.3
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0824	0.018	-4.480	0.000	-0.119	-0.046
horsepower	0.4402	0.052	8.390	0.000	0.336	0.544
carwidth	0.3957	0.046	8.677	0.000	0.306	0.486
hatchback	-0.0414	0.013	-3.219	0.002	-0.067	-0.016
Highend	0.2794	0.022	12.591	0.000	0.236	0.323

Omnibus:	29.385	Durbin-Watson:	1.955
Prob(Omnibus):	0.000	Jarque-Bera (JB):	98.010
Skew:	0.692	Prob(JB):	5.22e-22
Kurtosis:	6.812	Cond. No.	12.9

Figure 9 : Final result of OLS Regression Model

The final model has a high R-squared value of 0.899, indicating that 89.9% of the variability in the dependent variable (price) is explained by the independent variables included in the model. The adjusted R-squared value of 0.896 suggests that the independent variables are good predictors of the dependent variable. The coefficients of the independent variables **'horsepower', 'carwidth', 'hatchback', and 'Highend'** are 0.4402, 0.3957, -0.0414, and 0.2794 respectively. This suggests that an increase in horsepower, carwidth, and Highend status results in an increase in the price of the car, while hatchback status is associated with a decrease in price. The p-values of all the independent variables are less than 0.05, which indicates that all variables are statistically significant in predicting the dependent variable. However, the high value of the Omnibus test statistic (29.385) suggests that the residuals may not be normally distributed, which could potentially impact the reliability of the model. Additionally, the Durbin-Watson test statistic of 1.955 suggests that there may be some autocorrelation present in the residuals.

Overall, the equation of price of a car in US could be written as:

Price= -0.0824 + 0.4402 * 'housepower' + 0.3957 * 'carwidth'

-0.0414 * 'hatchback' + 0.2794 * 'highend'

4 Results

After establishing the final model, I made an experiment to evaluate the performance of the final model on the test data set, which was not used during the model training process. The model is then used to predict the target variable on the test data set, and the R-squared value is computed to measure how well the model fits the test data. The R-squared score of the final model is about 0.8614, which could be considered a good model performance. Then I plot the error distribution figure and test data and train data scatter according to the results of the experiment.

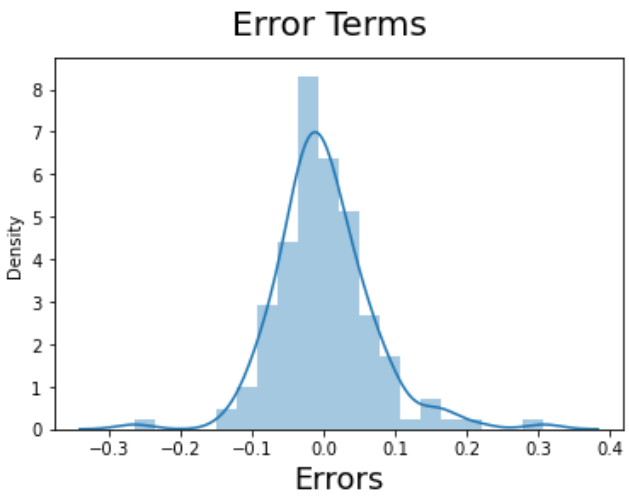

Figure 10: Error terms of final model

From the Figure 10, we can find that the peak of the distribution is located around 0, which indicates that the model is unbiased, and the error terms are centered around the mean. The spread of the distribution is relatively narrow, indicating that the model is accurate and consistent, and the error terms are small.

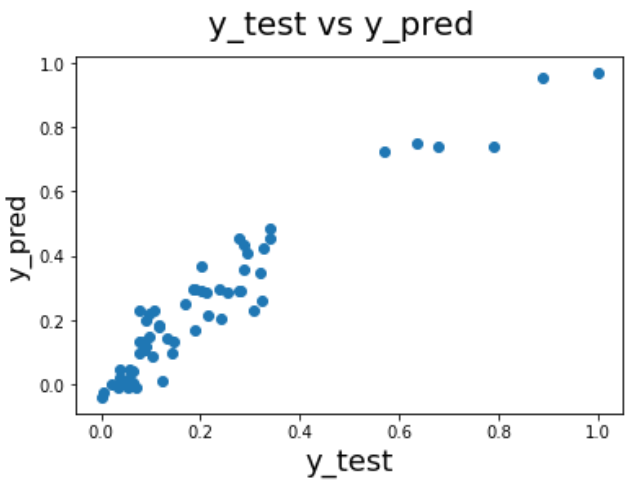

Figure 11: Evaluation of the model

From the Figure 11 we can find that the scatter plot is close to a diagonal line (i.e., the line with slope 1), it indicates that the model predictions are close to the actual values, and the final model is performing well.

5 Conclusions

In conclusion, this paper analyzed the variables affecting car price prices in the US, using descriptive statistics and regression analysis. Using variable selection techniques such as VIF and OLS, we developed a regression model with an R-squared score of 0.8614, indicating a good fit between the model and the data. The most important variables that influence the price of a car were found to be **'horsepower', 'carwidth','hatchback'(**Representing a kind of **'carbody'), and 'highend'** (Representing luxury car brands, such as Porsche, BMW, etc.).

The variables **'horsepower', 'carwidth', and 'highend'** are positively correlated with the **'price'** variable, meaning that an increase in these variables is associated with an increase in the **'price'**. On the other hand, the **'hatchback'** variable is negatively correlated with the **'price'**, meaning that an increase in the 'hatch' variable is associated with a decrease in the 'price'. the equation of price of a car in US could be written as:

Price= -0.0824 + 0.4402 * 'housepower' + 0.3957 * 'carwidth'- 0.0414 * 'hatchback' + 0.2794 * 'highend'

However, there is still room for further work. Advanced modeling techniques, such as Random Forest and Gradient Boosting, could be employed to improve the model's performance and predictive accuracy. Additionally, the dataset could be expanded to include more variables, such as fuel efficiency, safety ratings. What's more, the dataset only has 205 observations so I can't train a good enough model. With a large dataset, there are several opportunities for further work to enhance the accuracy and interpretability of the model.

Overall, this study provides valuable insights into the variables that affect car prices in the US and offers potential avenues for future research to enhance our understanding of this complex and dynamic market.

REFERENCES

[1] Abdulsalam, Sulaiman Olaniyi, et al. "Performance evaluation of ANOVA and RFE algorithms for classifying microarray dataset using SVM." *Information Systems: 17th European, Mediterranean, and Middle Eastern Conference, EMCIS 2020, Dubai, United Arab Emirates, November 25–26, 2020, Proceedings 17*. Springer International Publishing, 2020.

[2] Hayes, Andrew F., and Li Cai. "Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation." *Behavior research methods* 39 (2007): 709-722.

[3] Miles, Jeremy. "Tolerance and variance inflation factor." *Wiley statsref: statistics reference online* (2014).