

4KG3 Assignment 1 – Linear regression for prediction

Student Name: Jieyu Zhang

Student Number: 400368960

Objectives: Learn how to use linear regression for prediction.

1. Using data visualization for initial variable investigation and selection

(a) What are the distributions for the dependent and independent variables?

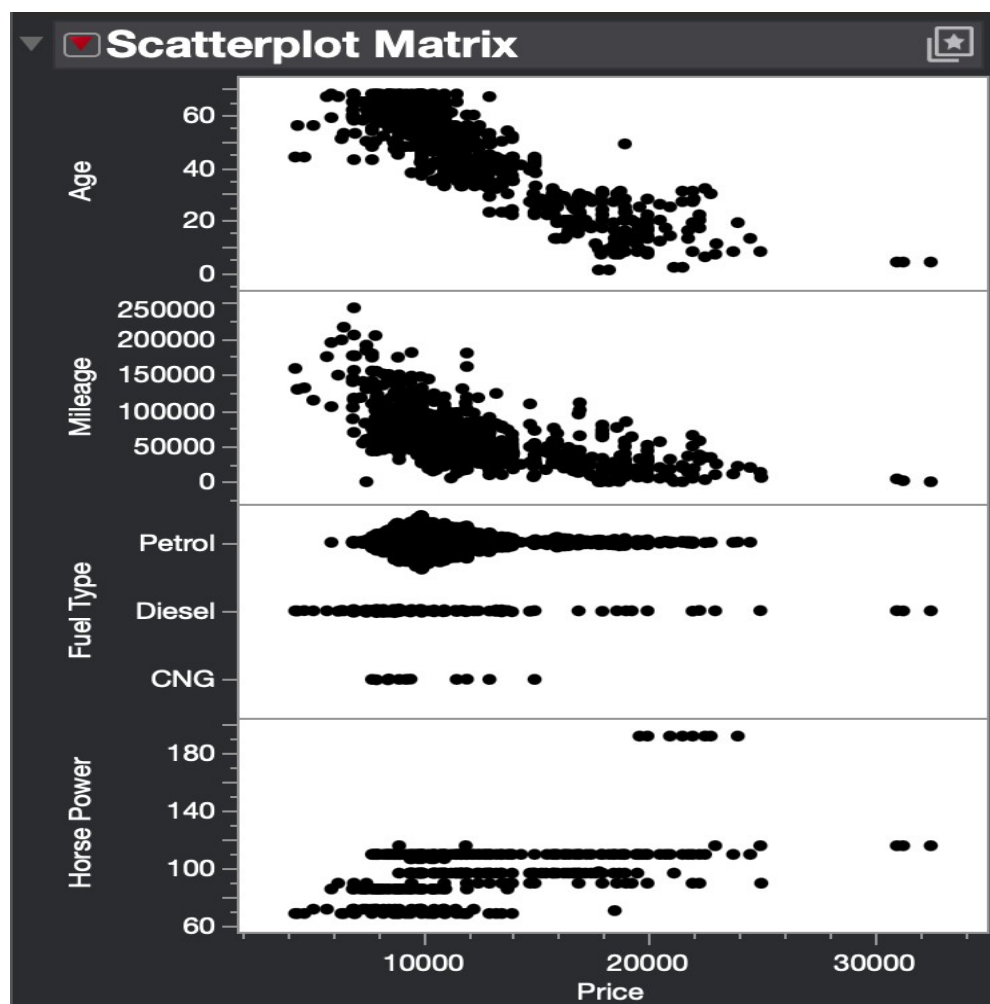
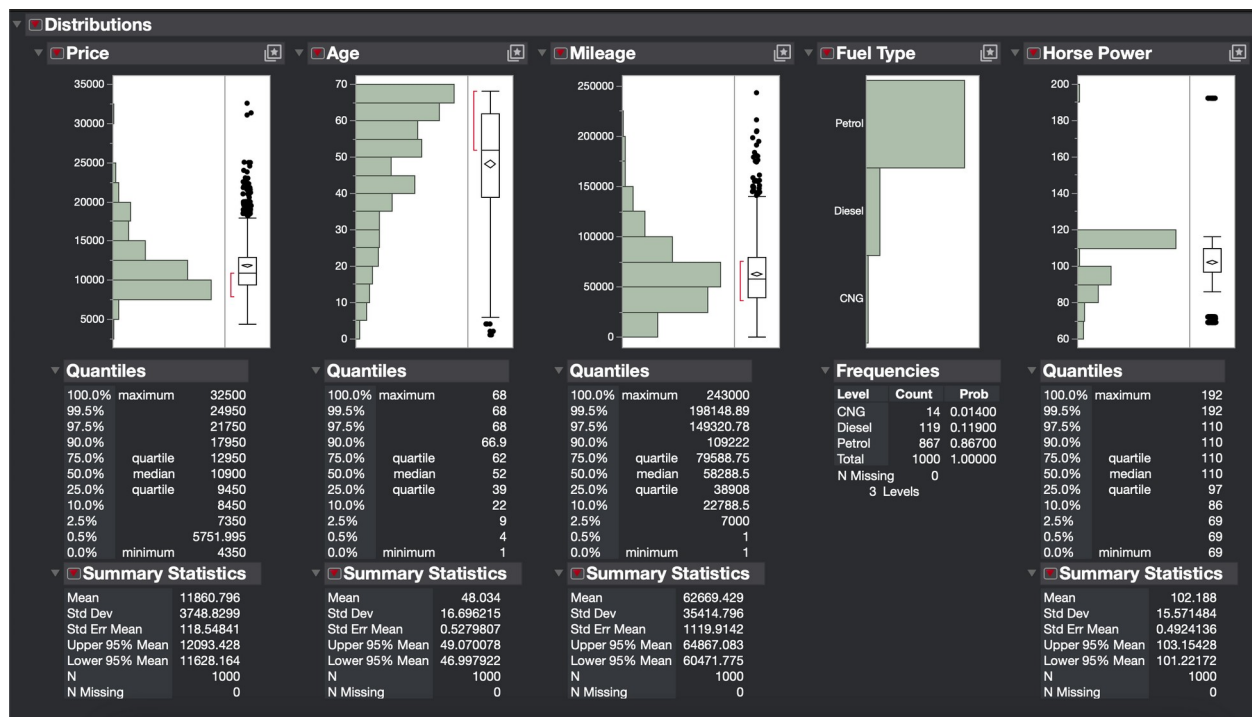
Price, the dependent variable, is right-skewed, with a number of outliers over 25,000 and the majority of values concentrated below 15,000. While mileage is right-skewed, meaning most cars have lesser mileage but some have more than 150,000, age is left-skewed, suggesting there are more older automobiles. Fuel type is classified, with gasoline accounting for the majority (86.7%), followed by diesel (11.9%) and compressed natural gas (1.4%). With a few high-value outliers reaching 192 HP, the horse power is roughly typical, centered at 110 HP.

(b) How do they related to each other?

Older automobiles and autos with more miles on them often cost less since the dependent variable, price, has a negative relationship with both age and mileage. Since automobiles with greater horsepower often cost more, there is a positive link between horsepower and price. The most popular fuel is gasoline. Mileage and Horse Power outliers point to possible non-linear correlations.

(c) What appears to be the three or four most important car specifications for predicting the used car price?

When estimating the price of a secondhand automobile, age, mileage, horsepower, and fuel type are the most crucial factors. Older automobiles and cars with more miles on them are often less expensive since age and mileage have an inverse relationship with pricing. Given that higher-performance automobiles tend to fetch higher prices, horsepower and price seem to be positively connected. Fuel type also matters since market demand and efficiency may influence prices for various fuels (Petrol, Diesel, and CNG). These four variables are important predictors of auto pricing since they show a great deal of variation.



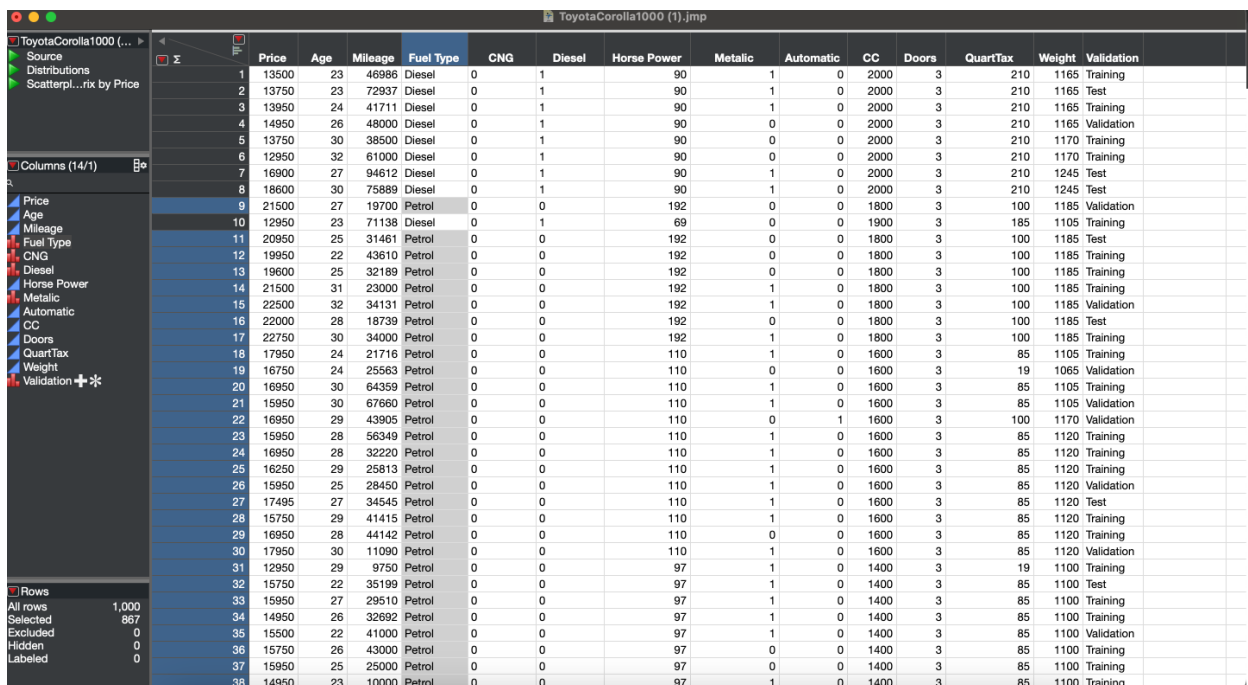
2. Data preparation and partition for training, validation, and testing

(a) Why do we need to convert fuel type to dummy variables?

Since Fuel Type is a categorical variable and the majority of statistical and machine learning models demand numerical input, we must turn it into dummy variables. To enable models to grasp how fuel type affects pricing, we create dummy variables that convert categories (such as gasoline, diesel, and CNG) into binary values (0 or 1). This guarantees accurate comparison between fuel types and avoids erroneous assumptions about numerical correlations.

(b) Why do we need to do data partition?

Data partitioning is necessary to prevent overfitting and to guarantee that machine learning models generalize effectively to new data. We enable the model to learn patterns from one part of the dataset (training set), adjust hyperparameters using another (validation set), and assess performance on entirely unknown data (test set) by splitting the dataset into training, validation, and test sets. By distributing the dependent variable (price) evenly among these sets, stratified partitioning preserves uniformity and guards against bias.



	Price	Age	Mileage	Fuel Type	CNG	Diesel	Horse Power	Metallic	Automatic	CC	Doors	QuartTax	Weight	Validation
1	13500	23	46986	Diesel	0	1	90	1	0	2000	3	210	1165	Training
2	13750	23	72937	Diesel	0	1	90	1	0	2000	3	210	1165	Test
3	13950	24	41711	Diesel	0	1	90	1	0	2000	3	210	1165	Training
4	14950	26	48000	Diesel	0	1	90	0	0	2000	3	210	1165	Validation
5	13750	30	38500	Diesel	0	1	90	0	0	2000	3	210	1170	Training
6	12950	32	61000	Diesel	0	1	90	0	0	2000	3	210	1170	Training
7	16900	27	94612	Diesel	0	1	90	1	0	2000	3	210	1245	Test
8	18600	30	75889	Diesel	0	1	90	1	0	2000	3	210	1245	Test
9	21500	27	19700	Petrol	0	0	192	0	0	1800	3	100	1185	Validation
10	12950	23	71138	Diesel	0	1	69	0	0	1900	3	185	1105	Training
11	20950	25	31461	Petrol	0	0	192	0	0	1800	3	100	1185	Test
12	19950	22	43610	Petrol	0	0	192	0	0	1800	3	100	1185	Training
13	19600	25	32189	Petrol	0	0	192	0	0	1800	3	100	1185	Training
14	21500	31	23000	Petrol	0	0	192	1	0	1800	3	100	1185	Training
15	22500	32	34131	Petrol	0	0	192	1	0	1800	3	100	1185	Validation
16	22000	28	18739	Petrol	0	0	192	0	0	1800	3	100	1185	Test
17	22750	30	34000	Petrol	0	0	192	1	0	1800	3	100	1185	Training
18	17950	24	21716	Petrol	0	0	110	1	0	1600	3	85	1105	Training
19	16750	24	25563	Petrol	0	0	110	0	0	1600	3	19	1065	Validation
20	16950	30	64359	Petrol	0	0	110	1	0	1600	3	85	1105	Training
21	15950	30	67660	Petrol	0	0	110	1	0	1600	3	85	1105	Validation
22	16950	29	43905	Petrol	0	0	110	0	1	1600	3	100	1170	Validation
23	15950	28	56349	Petrol	0	0	110	1	0	1600	3	85	1120	Training
24	16950	28	32220	Petrol	0	0	110	1	0	1600	3	85	1120	Training
25	16250	29	25813	Petrol	0	0	110	1	0	1600	3	85	1120	Training
26	15950	25	28450	Petrol	0	0	110	1	0	1600	3	85	1120	Validation
27	17495	27	34545	Petrol	0	0	110	1	0	1600	3	85	1120	Test
28	15750	29	41415	Petrol	0	0	110	1	0	1600	3	85	1120	Training
29	16950	28	44142	Petrol	0	0	110	0	0	1600	3	85	1120	Training
30	17950	30	11090	Petrol	0	0	110	1	0	1600	3	85	1120	Validation
31	12950	29	9750	Petrol	0	0	97	1	0	1400	3	19	1100	Training
32	15750	22	35199	Petrol	0	0	97	1	0	1400	3	85	1100	Test
33	15950	27	29510	Petrol	0	0	97	1	0	1400	3	85	1100	Training
34	14950	26	32692	Petrol	0	0	97	1	0	1400	3	85	1100	Training
35	15500	22	41000	Petrol	0	0	97	1	0	1400	3	85	1100	Validation
36	15750	26	43000	Petrol	0	0	97	0	0	1400	3	85	1100	Training
37	15950	25	25000	Petrol	0	0	97	0	0	1400	3	85	1100	Training
38	14950	23	10000	Petrol	0	0	97	1	0	1400	3	85	1100	Training

3. Run a linear regression with all available variables

(a) What is the mathematical formula of the regression model obtained?

Price = -6452.144 + (-128.645 * Age) + (-0.019 * Mileage + (-2065.622 * Fuel Type[CNG] or 25.500 * Fuel Type[Diesel] or 2040.122 * Fuel Type[Petrol]) + 46.341 * Horse Power + (-

1.782 * Metallic(0) or 1.782 * Metallic(1)) + 465.234 * Automatic + (-2.009 * CC) + (-87.320 * Doors + 24.562 * QuartTax + 19.014 * Weight

$$\begin{aligned}
 & -6452.144443 \\
 & + -128.6449702 \cdot \text{Age} \\
 & + -0.019014522 \cdot \text{Mileage} \\
 & + \text{Match}(\text{Fuel Type}) \begin{pmatrix} \text{"CNG"} \Rightarrow -2065.622469 \\ \text{"Diesel"} \Rightarrow 25.500412656 \\ \text{"Petrol"} \Rightarrow 2040.1220566 \\ \text{else} \Rightarrow . \end{pmatrix} \\
 & + 46.34083409 \cdot \text{Horse Power}
 \end{aligned}$$

$$\begin{aligned}
 & + \text{Match}(\text{Metalic}) \begin{pmatrix} 0 \Rightarrow -1.781987103 \\ 1 \Rightarrow 1.781987103 \\ \text{else} \Rightarrow . \end{pmatrix} \\
 & + 465.23358427 \cdot \text{Automatic} \\
 & + -2.00940119 \cdot \text{CC} \\
 & + -87.32045424 \cdot \text{Doors} \\
 & + 24.561784532 \cdot \text{QuartTax} \\
 & + 19.013524875 \cdot \text{Weight}
 \end{aligned}$$

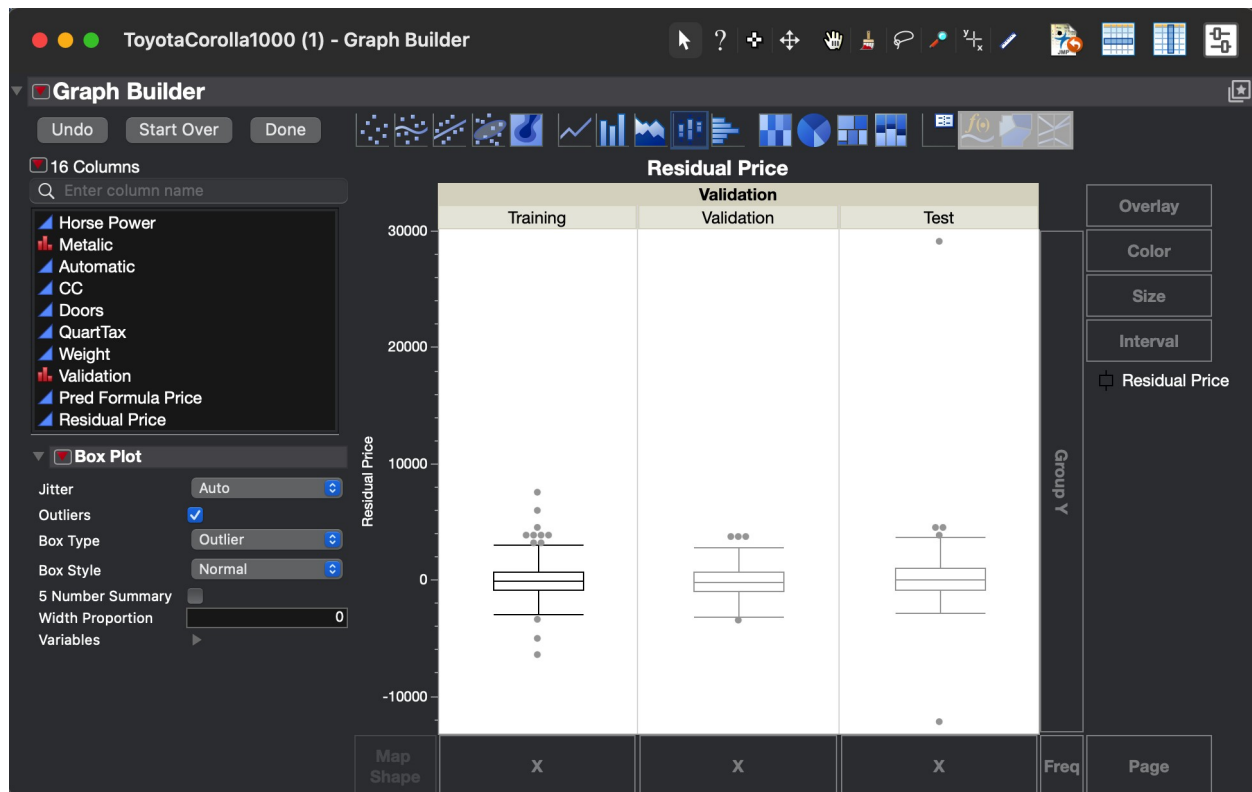
(b) How do you calculate the predicted price and the prediction error (residual) for each record?

To calculate the predicted price we just have to put the value of Age, Mileage, Fuel Type, Horse Power, Metalic, Automatic, CC, Doors, QuartTax and Weight in to the formula that is showing in question (a).

To calculate the Residual we can use “Residual = Actual Price – Predicted Price”, we just have to put the value in the “Price” column as Actual Price and value of “Pred Formula Price” column as Predicted Price.

- (c) Show the error distribution for training, validation and testing? What are the differences between them?

The training set having the fewest outliers and the least residual variance. Although the residuals in the validation set vary a little more, they are still comparable to those in the training set, indicating good generalization. The test set, on the other hand, includes the most severe outliers and the highest residual spread, indicating that the model may be overfitting to the training set and struggles with unknown data.



4. Automated variable selection

- (a) What methods have you used for variable selection?

For variable selection, the stepwise regression method was utilized, especially the forward selection approach with the Maximum Validation R-Square stopping criteria. This strategy begins with no variables and gradually introduces predictors depending on their statistical significance and contribution to improved model performance.

- (b) What is the best set of variables you will use? What are the criteria used for selection?

Age, Mileage, Horse Power, CC (Engine Size), Doors, QuartTax, and Weight are the most effective variables for forecasting automobile prices, with p-values (<0.05) indicating statistical significance. These variables were chosen by stepwise regression (forward selection) with Maximum Validation R-Square as the stopping criteria. The selection criteria include p-values to assure significance, F-Ratio to quantify contribution, Adjusted R-Square for model fit, and AIC/BIC to balance complexity and accuracy. Less relevant variables, such as fuel type, automatic transmission, and metallic paint, were eliminated to avoid overfitting and enhance model efficiency.

ToyotaCorolla1000 (1) - Fit Stepwise

Stepwise Fit for Price

Stepwise Regression Control

Stopping Rule: Max Validation RSquare → Enter All Make Model

Direction: Forward ← Remove All Run Model

Rules: Combine

Go Stop Step

Training Rows 500
Validation Rows 300
Test Rows 200

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC	RSquare Validation	RASE Validation	RSquareTest	RASE Test
979357403	492	1410.8734	0.8630	0.8611	55.803458	8	8681.205	8718.77	0.8979	1204.729	0.1699	3300.034

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☒	☒	Intercept	-2779.2009	1	0	0.000	1
☐	☒	Age	-127.35832	1	1.12e+9	562.406	1.8e-83
☐	☒	Mileage	-0.0219574	1	1.654e+8	83.084	2e-18
☐	☐	CNG{1-0}	0	1	76332182	41.504	2.8e-10
☐	☐	Diesel{0-1}	0	1	351284.8	0.176	0.67486
☐	☒	Horse Power	54.8433397	1	2.912e+8	146.268	1.2e-29
☐	☐	Metallic{0-1}	0	1	72891.78	0.037	0.84847
☐	☐	Automatic	0	1	1254928	0.630	0.42775
☐	☒	CC	-2.818425	1	53127568	26.690	3.48e-7
☐	☒	Doors	-38.937653	1	641565.2	0.322	0.57049
☐	☒	QuartTax	14.1099508	1	66093773	33.204	1.47e-8
☐	☒	Weight	18.4174857	1	1.985e+8	99.699	1.7e-21

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	RSquare Validation
1	Age	Entered	0.0000	5.179e+9	0.7242	590.88	2	9018.82	9031.41	0.7863
2	Age	Removed	0.0000	5.179e+9	0.0000	3443.5	1	9660.92	9669.33	-0.000
3	Diesel{0-1}	Entered	0.3794	11096561	0.0016	3439.4	2	9662.17	9674.77	0.0028
4	CNG{1-0}	Entered	0.1869	25007950	0.0050	3427.6	3	9662.45	9679.23	-0.001
5	Horse Power	Entered	0.0000	1.555e+9	0.2225	2572.4	4	9541.16	9562.11	0.1697
6	CNG{1-0}	Removed	0.0510	42896009	0.2165	2594	3	9542.96	9559.74	0.1735
7	Horse Power	Removed	0.0000	1.527e+9	0.0016	3439.4	2	9662.17	9674.77	0.0028

Step History										
Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	RSquare Validation
1	Age	Entered	0.0000	5.179e+9	0.7242	590.88	2	9018.82	9031.41	0.7863
2	Age	Removed	0.0000	5.179e+9	0.0000	3443.5	1	9660.92	9669.33	-0.000
3	Diesel{0-1}	Entered	0.3794	11096561	0.0016	3439.4	2	9662.17	9674.77	0.0028
4	CNG{1-0}	Entered	0.1869	25007950	0.0050	3427.6	3	9662.45	9679.23	-0.001
5	Horse Power	Entered	0.0000	1.555e+9	0.2225	2572.4	4	9541.16	9562.11	0.1697
6	CNG{1-0}	Removed	0.0510	42896009	0.2165	2594	3	9542.96	9559.74	0.1735
7	Horse Power	Removed	0.0000	1.537e+9	0.0016	3439.4	2	9662.17	9674.77	0.0028
8	Diesel{0-1}	Removed	0.3794	11096561	0.0000	3443.5	1	9660.92	9669.33	-0.000
9	Age	Entered	0.0000	5.179e+9	0.7242	590.88	2	9018.82	9031.41	0.7863
10	Mileage	Entered	0.0000	2.317e+8	0.7567	465.17	3	8958.35	8975.13	0.7931
11	Horse Power	Entered	0.0000	2.385e+8	0.7900	335.7	4	8886.68	8907.64	0.8180
12	Automatic	Entered	0.1767	5531308	0.7908	334.65	5	8886.89	8912.01	0.8194
13	CC	Entered	0.0000	1.059e+8	0.8056	278.3	6	8852.25	8881.53	0.8299
14	Doors	Entered	0.1456	5966000	0.8064	277.01	7	8852.17	8885.59	0.8311
15	QuartTax	Entered	0.0000	2.156e+8	0.8366	160.18	8	8769.59	8807.16	0.8592
16	Weight	Entered	0.0000	1.906e+8	0.8632	57.112	9	8682.65	8724.34	0.8976
17	Weight	Removed	0.0000	1.906e+8	0.8366	160.18	8	8769.59	8807.16	0.8592
18	QuartTax	Removed	0.0000	2.156e+8	0.8064	277.01	7	8852.17	8885.59	0.8311
19	CC	Removed	0.0000	1.034e+8	0.7920	332.03	6	8886.13	8915.41	0.8215
20	Automatic	Removed	0.1741	5579384	0.7912	333.1	5	8885.95	8911.06	0.8201
21	Mileage	Removed	0.0000	90413358	0.7785	380.94	4	8913.29	8934.24	0.8177
22	Age	Removed	0.0000	4.472e+9	0.1531	2843.9	3	9581.87	9598.64	0.1518
23	Horse Power	Removed	0.0000	8.939e+8	0.0281	3334.6	2	9648.67	9661.27	0.0427
24	Doors	Removed	0.0002	2.013e+8	0.0000	3443.5	1	9660.92	9669.33	-0.000
25	Automatic	Removed	0.4278	1254928	0.8630	55.803	8	8681.21	8718.77	0.8979

5. Possible real word applications

- (a) What is the potential business value for used car price prediction? For instance, you may consider its possible use for used car sellers or replacement cost evaluation for insurance companies.

Many companies benefit from the ability to estimate used vehicle prices. Used car dealers may establish fair and competitive rates based on characteristics such as age, mileage, and vehicle attributes, allowing them to sell automobiles faster and for a higher profit. Insurance firms may utilize price projections to determine the appropriate reimbursement when a vehicle is damaged or lost, making claims more equitable. Banks and lenders can use it to determine loan amounts by estimating how much an automobile will be valued in the future. Car buyers and dealerships also profit since it allows them to determine whether a price is appropriate, making purchasing and selling cars easier and more transparent.

- (b) Search the web and find out potential utilities of linear regression in other domains such as finance, healthcare etc.

One use of linear regression in finance is the Capital Asset Pricing Model (CAPM), which assists investors in determining an asset's expected return depending on its risk. The formula: $\text{Expected return} = \text{Risk-Free Rate} + \text{Beta} * (\text{Market return} - \text{Risk-Free Rate})$ use linear regression to estimate beta, a measure of an asset's volatility compared to the market. This allows investors to determine if an asset provides a good return for its degree of risk.

6. Report how much time you have spent on this assignment, what problem you have faced and what you have learned.

I worked on this project for around 5 hours, which included data analysis, model selection, and interpretation. One of the most difficult issues was dealing with variable selection, namely selecting the most important predictors while avoiding overfitting. Furthermore, understanding statistical outputs like p-values, R-squared, and residual distributions needed considerable consideration. Another challenge was understanding the influence of category factors and transforming them into useful dummy variables.

This project taught me how to utilize stepwise regression for feature selection, residual analysis to measure model performance, and several assessment metrics such as AIC, BIC, and RASE. I also learned more about the commercial benefits of predictive modeling, notably in pricing strategies for used automobiles and insurance applications. This experience helped me improve my abilities in data-driven decision-making and statistical modeling.