Linear Regression: Car Price Prediction

1. Description of your dataset: resource, dimension, variable description, etc.

We acquired a kaggle data set named Car Price Prediction Multiple Linear Regression with 11 categorical variables and 15 numeric variables, while including car_ID and our response variable, which is the price of the car.

Our **categorical variables**, excluding CarName, and their **distinct values** include:

```
1: symboling: [ 3  1  2  0 -1 -2]
2: fueltype: ['gas' 'diesel']
3: aspiration: ['std' 'turbo']
4: doornumber: ['two' 'four']
5: carbody: ['convertible' 'hatchback' 'sedan' 'wagon' 'hardtop']
6: drivewheel: ['rwd' 'fwd' '4wd']
7: enginelocation: ['front' 'rear']
8: enginetype: ['dohc' 'ohcv' 'ohc' 'l' 'rotor' 'ohcf' 'dohcv']
9: cylindernumber: ['four' 'six' 'five' 'three' 'twelve' 'two' 'eight']
10: fuelsystem: ['mpfi' '2bbl' 'mfi' '1bbl' 'spfi' '4bbl' 'idi' 'spdi']
```

Our **numeric variables** and their **ranges** include:

```
1: car_ID: 1 - 205
2: wheelbase: 86.6 - 120.9
3: carlength: 141.1 - 208.1
4: carwidth: 60.3 - 72.3
5: carheight: 47.8 - 59.8
6: curbweight: 1488 - 4066
7: enginesize: 61 - 326
8: boreratio: 2.54 - 3.94
9: stroke: 2.07 - 4.17
10: compressionratio: 7.0 - 23.0
11: horsepower: 48 - 288
12: peakrpm: 4150 - 6600
13: citympg: 13 - 49
14: highwaympg: 16 - 54
15: price: 5118.0 - 45400.0
```

The prices of the vehicles ranged from $5,118.00, being the cheapest vehicle, to $45,400.00, being the most expensive. Overall there are 26 variables and 205 observations in the data set. Car ID is a primary, meaning each observation has a unique value. Hence Car ID is definitely not relevant for our linear regression analysis of the data. Our variables refer to different car metrics to extract data regarding car composition, size, speed and durability.

2. Statement of the research

In our project we were tasked to see which variables affect car price, how they affect price, and to what extent they affect price. The overall goal is to model the price of cars based on the independent variables. We used ordinary least squares regression to regress price against several variables. This method allows us to minimize square errors, estimating coefficients of linear regression equations that describe the relationship between the independent variables and a dependent variable (simple or multiple linear regression). We included dummy variables so that we were able to incorporate categorical variables in our analysis, alongside the quantitative values. We then trimmed this model to what we considered the best after doing several analyses on whether certain variables were significantly improving our model or simply causing overfitting or overcomplication. The steps of our analysis are as follows:

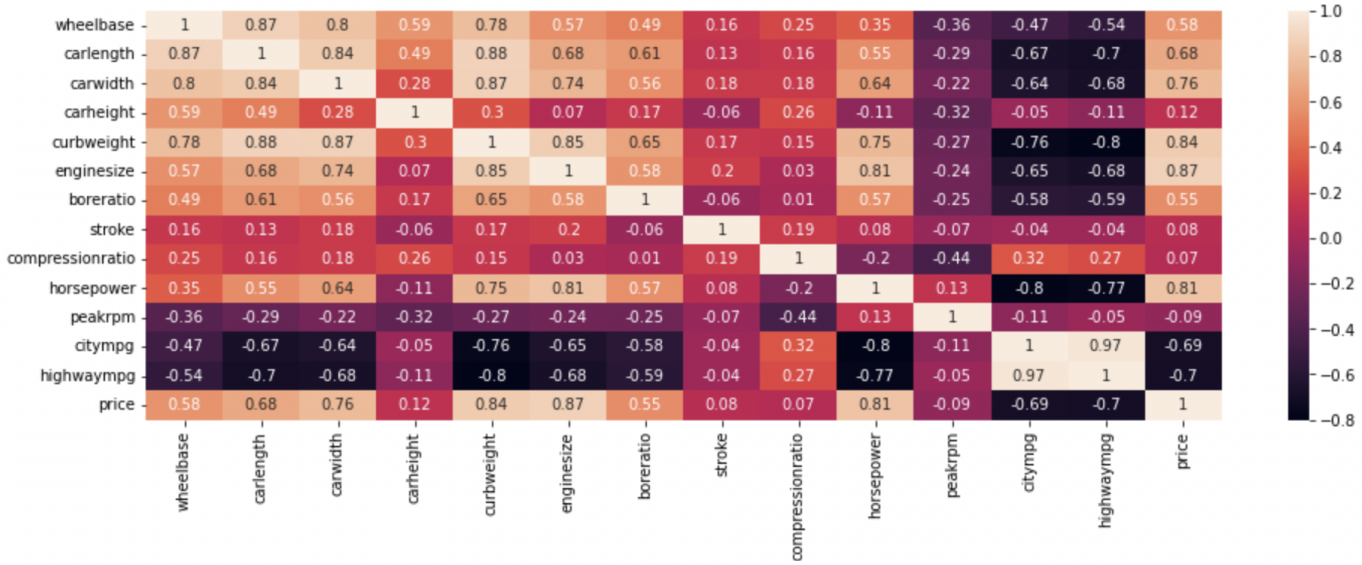
I. **Data Preprocessing** (loading the data, splitting into a testing and training, and exploratory data analysis)
II. **Fitting the full model**, then conducting ANOVA type 1,2, and 3.
III. **Model Diagnosis** (troubleshooting observations that violate the assumptions of linear regression)
    A. Heteroskedasticity
    B. Normality of residuals
    C. Multicollinearity
IV. **Model Selection** using BIC stepwise model selection
V. **Summarizing** results found

3. Data Preprocessing/ EDA

We split the data into training and validation data frames so we could validate the accuracy of our models. Because we were only provided 205 observations initially we did a 90-10 split, randomly placing 90 percent of our observations in the training data frame and 10 percent in the validation data frame. After splitting our training data frame had 184 observations, whereas the validation data frame had 21 observations.

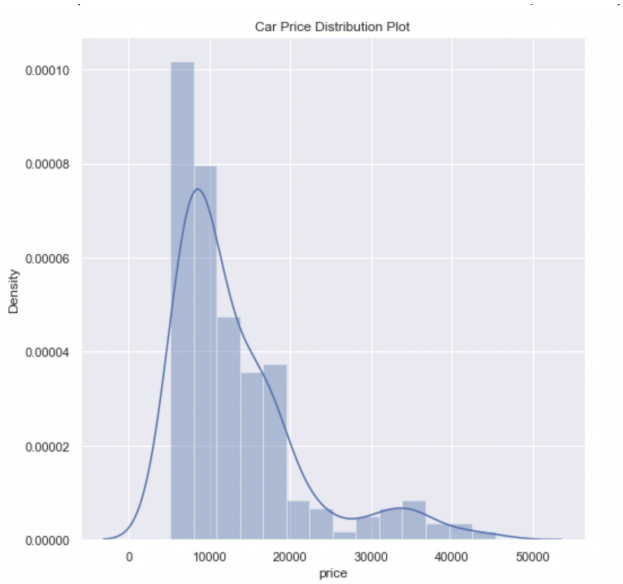
There were no null values in any of our variables in the dataset, so we did not need to delete or treat any observations because they were missing certain features.
We furthered our exploratory data analysis with graphs to dive deeper into the data and see factors such as correlations and distributions. Below we have a heat map illustrating correlation between different coefficients and a histogram of the car prices.

- Heatmap:



The heat map shows that there are several independent variables strongly correlated with price. From the graphic we see engine size has the highest positive correlation with price, with a rho value of 0.87 on a scale from -1 to 1. Similarly we can notice highway mpg has the strongest negative correlation to price, with a rho value of -0.7 on the same scale. There are 4 independent variables with a rho greater than 0.7 when analyzing their correlation with price. These variables are: horsepower, enginesize, curb weight and car width. We can also notice that there are several independent variables that are highly correlated with other independent variables. For instance, car width and car length are highly correlated, with a rho value of 0.84. This can cause multicollinearity within the regression analysis and in-turn inflate VIF scores. Keep in mind that the heat map does not take categorical data into consideration.

- Histogram of Price:



This histogram clearly demonstrates the the car prices are highly right skewed. Most of the cars fall between the price range $5,000.00 and $20,000.00. More than 15% of the cars are priced under $10,000. Limited cars cost more than $40,000.00. Less than 2% of the cars are priced over $40,000.00. The average price of the cars is $13276.71.

4. Model fitting

Initially we fit the full model using 23 predictors and 1 response variable, price. We excluded Car ID and Car name from this model because they were not relevant for our linear regression analysis. We used an ordinary least squares method, or OLS, meaning we found coefficients, β, by minimizing the error of the prediction, or the distance between the actual value and the predicted value. Some of the paramount statistics include:

- Adjusted R squared: 0.930
- BIC: 3502
- Prob (F-statistic): 5.01e-70

The adjusted R squared value measures the goodness of fit of the model, or in other words how well our model predicts the observations. Adjusted R squared is a better parameter than ordinary R squared when using multiple linear regression because it tests and considers different independent variables against the model. The adjusted R squared value ranges from 0 to 1, so an adjusted R squared value of 0.930 is good and means our data is fitting the model well.

BIC stands for Bayesian Information Criterion. Similar to the adjusted R squared value it also is a metric that measures the goodness of fit of a model. As the likelihood increases , the BIC decreases and as the complexity of the model increases the BIC also increases. Hence, we need to look for the model with the lowest BIC, meaning least complexity, but most likelihood.

To continue using a full model, in comparison to the intercept only model, we need to ensure that our model is passing the global F test. The P-value of our F-statistic is 5.01e-70, which is below .05 so we know that at least one of the predictors makes the model significantly better than the intercept only model.

- **ANOVA**

To ensure we are only choosing predictors which significantly better the model, we used ANOVA type 1, 2, and 3 hypothesis testing. ANOVA type 1 is sequential ANOVA, so the order of the predictors will affect the significance of the predictors. ANOVA type 2 is partial ANOVA so the significance result will be the same as t-test result in the summary table and can be different from the ANOVA type 1 table. ANOVA type 3 displays the same results as ANOVA type 2.

According to the ANOVA type 2 and and type 3 the significant predictors for the fitted model were [carbody, carwidth, curbweight, cylindernumber, enginelocation, enginesize, enginetype, fueltype, peakrpm, stroke] . According to ANOVA type 1 the significant predictors were aspiration, [carbody, carlength, carwidth, curbweight, cylindernumber, doornumber, drivewheel, enginelocation, enginesize, enginetype, fuelsystem, fueltype, horsepower, stroke, symboling, wheelbase]. After taking the intersection of the ANOVA type 1 and ANOVA type 2-3 results we get that the final predictors are: **[carbody, carlength carwidth, cylindernumber, enginelocation, enginesize, enginetype, fueltype, stroke]**.

5. Model Diagnosis

- ### Check constant variance

One of the key assumptions of linear regression is that the residuals are distributed with equal variance at each level of the predictor variable. This assumption is known as homoscedasticity.

When this assumption is violated, we say that heteroscedasticity is present in the residuals. When this occurs, the results of the regression become unreliable. We use Breusch-Pagan Test to determine if heteroscedasticity is present in the residuals or not.

The test uses the following null and alternative hypotheses:

- **Null Hypothesis (H0)**: Homoscedasticity is present (the residuals are distributed with equal variance)
- **Alternative Hypothesis (HA)**: Heteroscedasticity is present (the residuals are not distributed with equal variance)

Initially when we run the Breusch-Pagan Test we get:

Model: $e_i^2 = \gamma_0 + \gamma_1 x_{i,1} + \ldots + \gamma_{p-1} x_{i,p-1} + \xi_i$
$H_0 : \gamma_1 + \gamma_2 + \gamma_3 + \ldots + \gamma_{p-1} = 0$
H_1 : at least one $\gamma_k \neq 0$
$p\_value = 0.02646033548618978$
As $p\_value < \alpha$, indicates **significant heteroscendasticity problem**

To troubleshoot the heteroscedasticity problem we first see if deleting influential points and outliers provides a solution. We used two methods two identify outliers, Cook's distance and studentized residuals method. According to the Cook's distance method we have 23 candidate outlier observations and according to the studentized residuals method we have 11 candidate outlier observations. We take the intersection of these two sets and delete the influential points. We delete observations with the index { 2, 134, 167, 136, 137, 74, 16, 179, 180 }. In total we deleted 9 observations.

Without the influential points, we run the Breusch-Pagan Test again and we get:

Model: $e_i^2 = \gamma_0 + \gamma_1 x_{i,1} + \ldots + \gamma_{p-1} x_{i,p-1} + \xi_i$

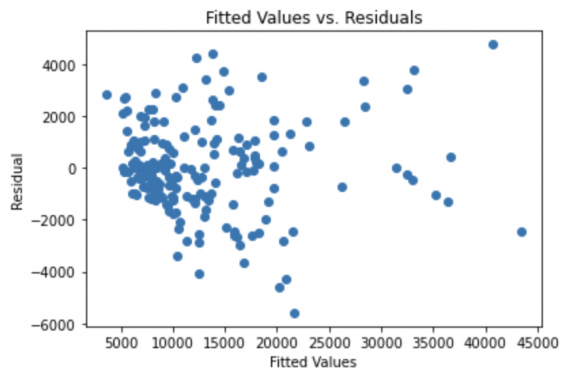
$H_0 : \gamma_1 + \gamma_2 + \gamma_3 + \ldots + \gamma_{p-1} = 0$

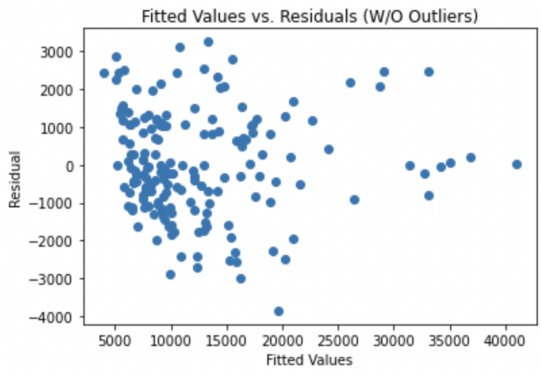
H_1 : at least one $\gamma_k \neq 0$

After removing the influential points, the $p\_value = 0.5459344719630755$.

As $p\_value > \alpha$, indicates **no heteroscendasticity problem**

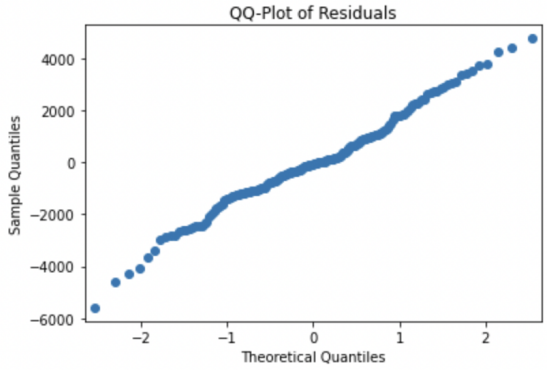
In this case we no longer have a heteroscedasticity problem. Hence we can conclude that removing the influential points successfully removed heteroscedasticity from the model. We see from the following residual plots, with and without the influential points, that the residuals are more evenly distributed at each level of the predicted variable after deleting influential points.

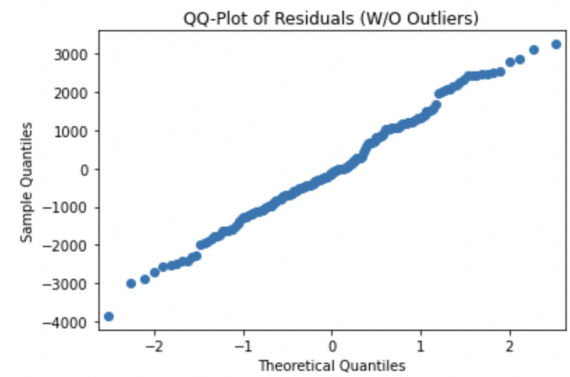




- **Check Normality of Residuals**

Another assumption is that the residuals follow a normal distribution. We can check this by plotting the residuals on a QQ plot. The straighter the line, the more normal our residuals are. Here are QQ plots with and without the influential points.





Both QQ plots show a straight line for the most part, but the values plotted on the right (in the graph without outliers) we clearly see a straighter line, meaning the residuals are distributed more normally without the outliers.

In addition to the QQ plots we also conducted the Jarque–Bera test to ensure sample data have the skewness and kurtosis matching a normal distribution. Our result after removing influential points is as follows:

> After removing influential point:
> $p\_value = 0.669$ for JB-Test
> $p\_value > \alpha$, **Fail to reject** H_0**,** there's **no significant** violation of normality

- **Check for Multicollinearity**

We did a check for multicollinearity using VIF scores, or variance inflation factor measure. High VIF scores means that the variable has high multicollinearity. In other words if a variable has a high VIF score it means that it is predicting the response variable in the same way as another independent variable, or the two independent variables are highly correlated with each other. Multicollinearity does not always affect the performance of the model, the adjusted R squared score, nor the accuracy of the model. For this reason we gave multicollinearity and VIF scores less weightage during model selection.

6. Model Selection

- **Stepwise BIC**

After all our analysis we chose to test all of our variables which passed at least one of the ANOVA tests (type 1, 2, or 3). We then added on a variable at a time starting with enginesize until we reached a point where the BIC was the lowest we could achieve. We obtained the lowest BIC with an early stop. The chosen predictors were [enginesize, drivewheel, cylindernumber, enginelocation, curbweight, enginetype, carwidth, carbody, aspiration, stroke, peakrpm, boreratio, wheelbase]. Then using the following table we are able to see that this final model, has a lower BIC than the full model, but still has a very similar adjusted R squared value.

	Models	Adj_RSquare	BIC	validate_err
0	Full Model(with inf_pt)	0.929671	3501.940736	11694.860673
1	Full Model(w/o inf_pt)	0.949584	3219.793604	10876.855658
2	Selected Model	0.949336	3150.480234	15143.586284

7. Summary and Findings

After removing the influence point, we fix the normality issue and keep the model away from heteroscedasticity. We use Step-BIC to find the best model using the updated dataset. The predictors we use in the final model include enginesize, drivewheel, cylindernumber, enginelocation, curbweight, englinetype, carwidth, carbody, aspiration, stroke, peakrpm, boreratio, and wheelbase. The final model with all the relevant predictors follows:

```
price ~ enginesize + drivewheel + cylindernumber + enginelocation + curbweight + enginetype + carwidth + carbody + aspiration + stroke + peakrpm + boreratio + wheelbase
```