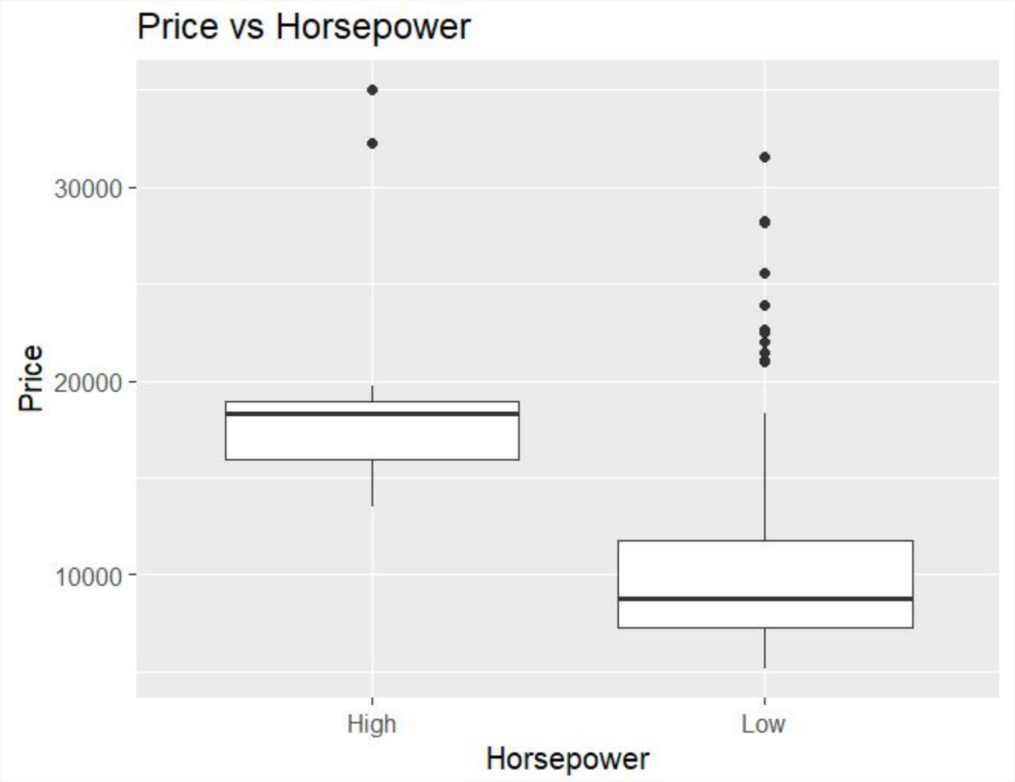
A Brief Exploration of the Automobile Market

The automobile market is dynamic and ever-changing, characterized by a wealth of options for consumers and constant technological improvements. This project uses hypothesis testing as part of a thorough investigation to identify important correlations and differences in the vehicle attributes.

Exploratory data analysis:

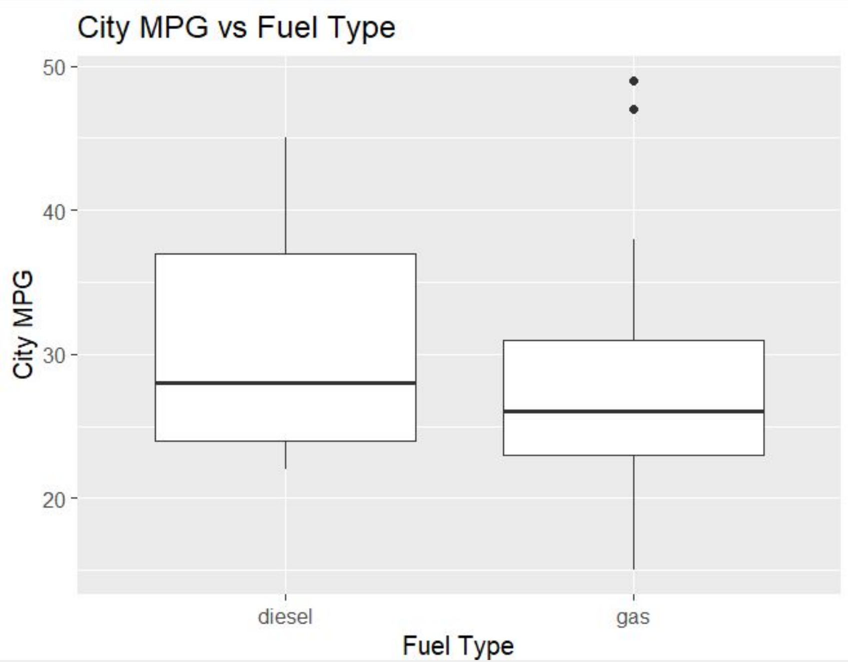
Box plots to understand the data and to understand how the horsepower is related to price.

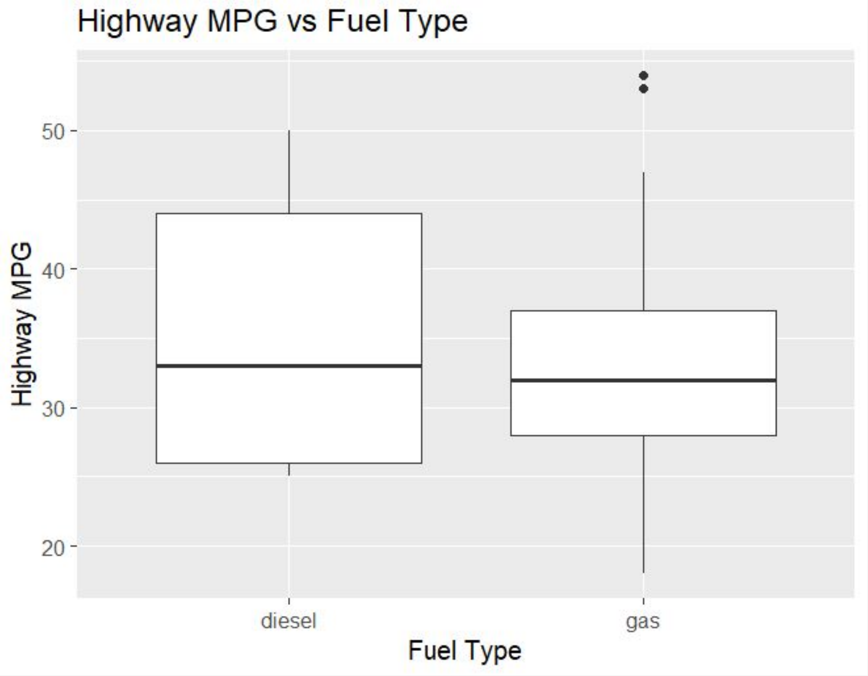
This was our inspiration for first hypothesis.



Exploratory data analysis(Continued):

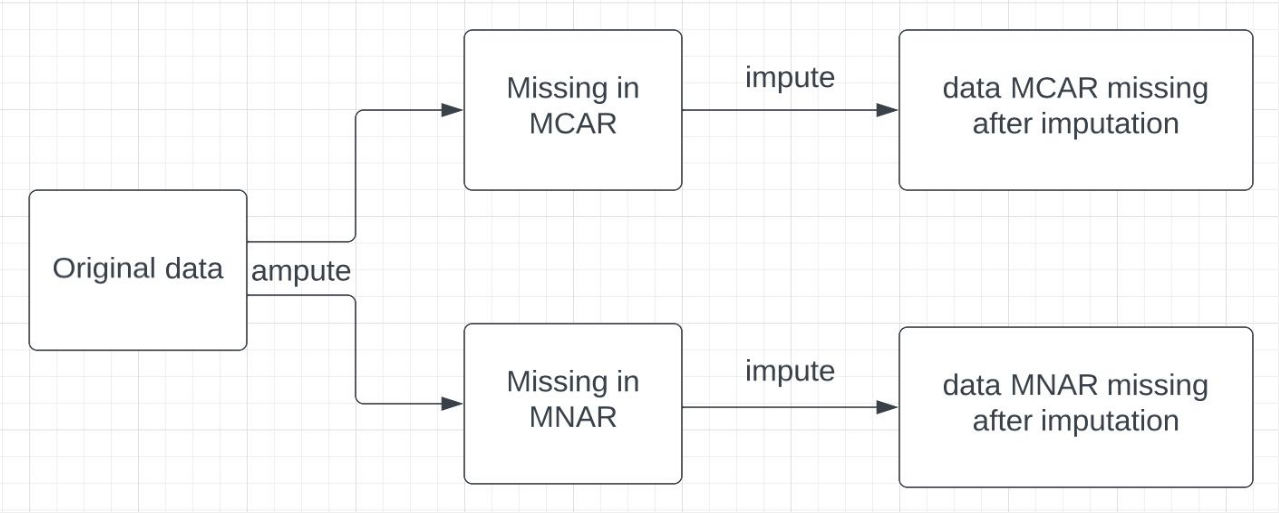
- Boxplot of city mpg vs fuel type and highway mpg vs fuel type. To check if there are any other underlying trends.



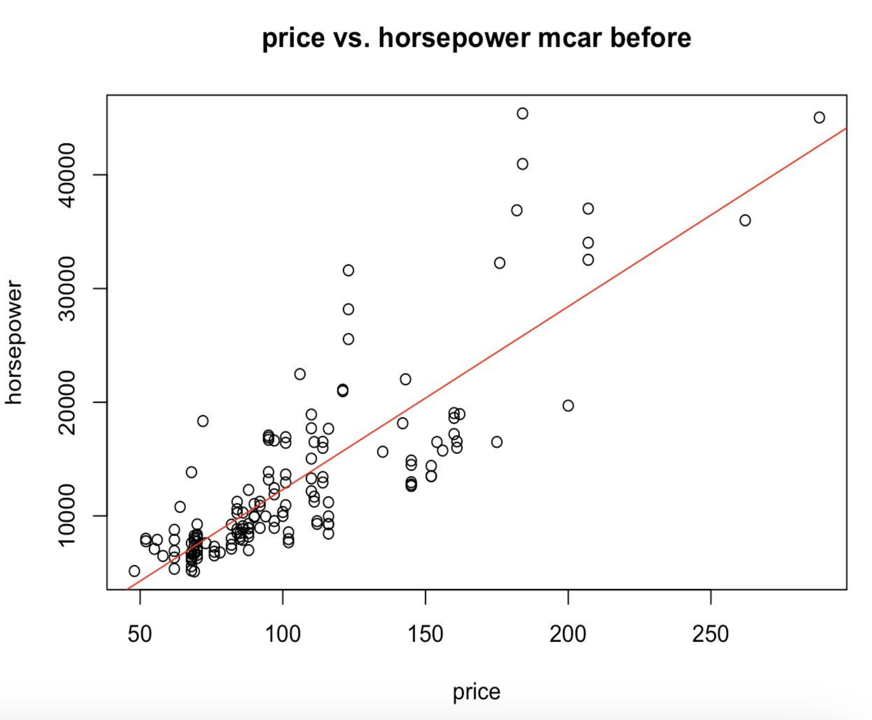


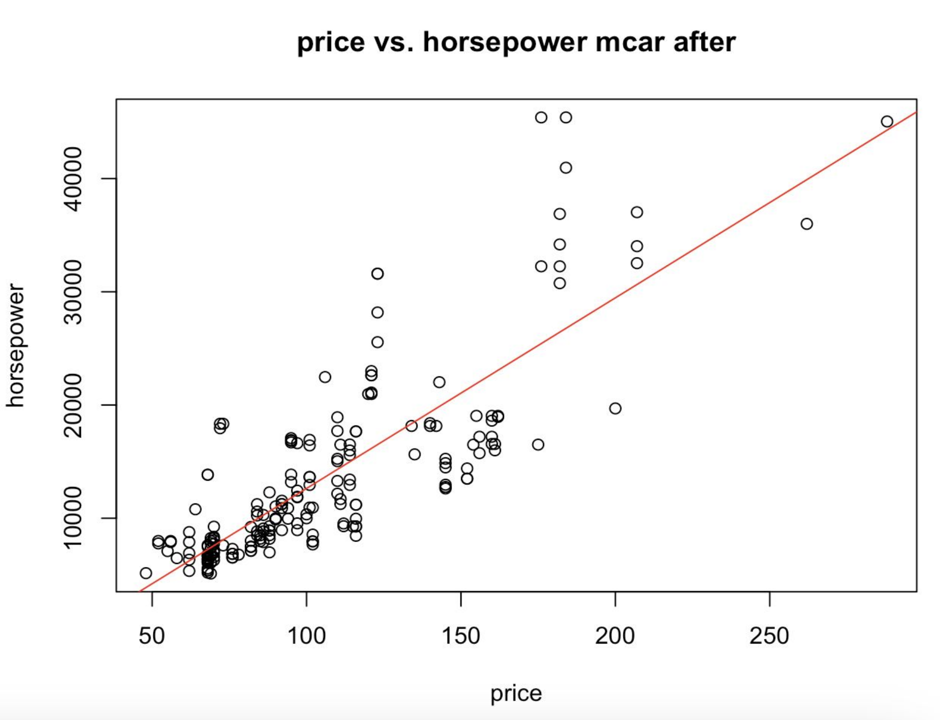
Handling Missing values:

- MCAR(Missing completely at random) type Missing
- MNAR(Missing not at random) type Missing

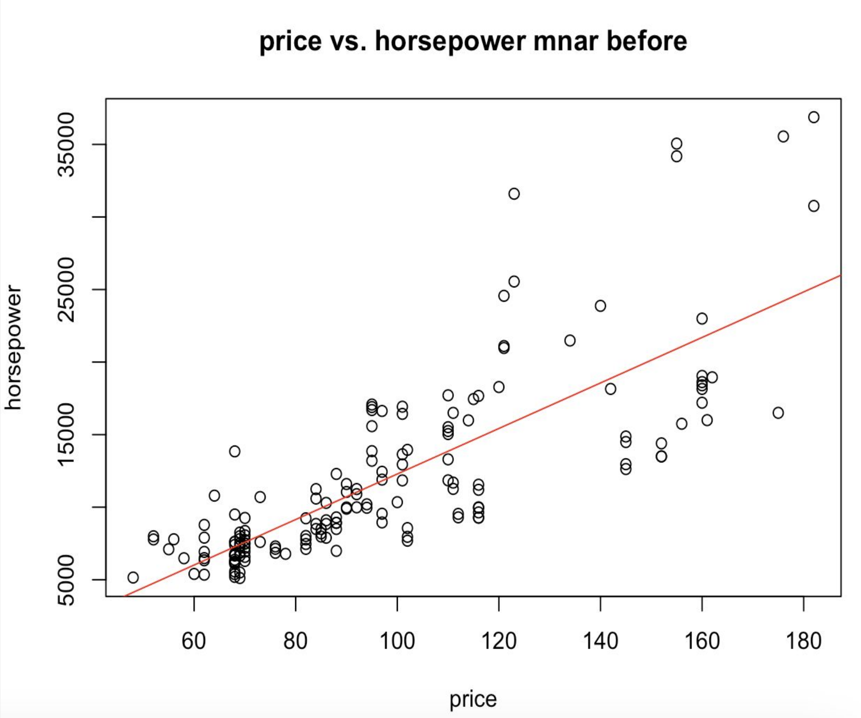

Original data
ampute
Missing in MCAR
Missing in MNAR
impute
impute
data MCAR missing after imputation
data MNAR missing after imputation

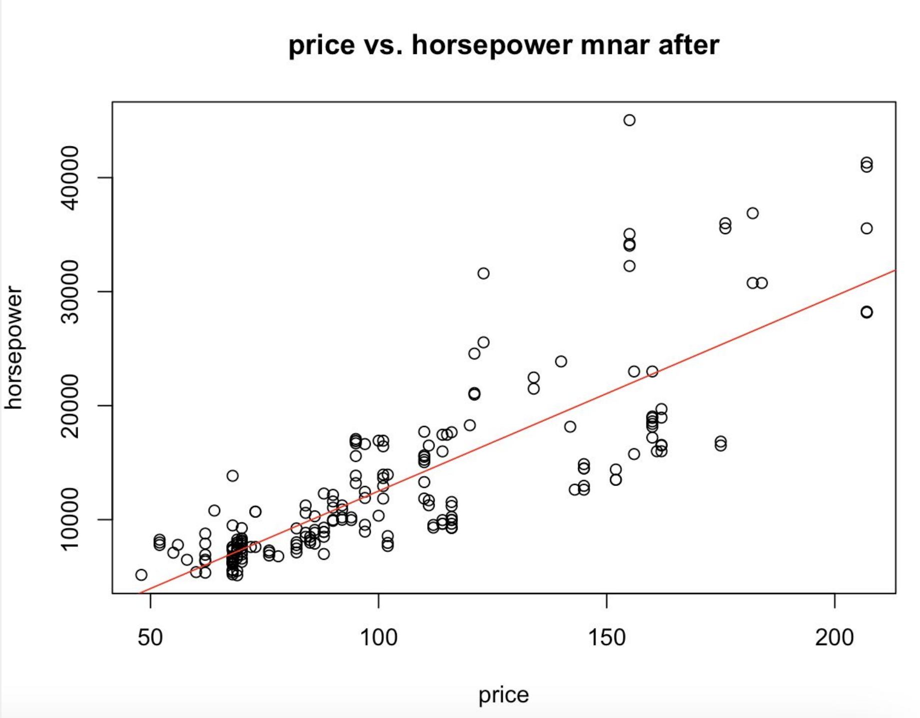
Plots on Imputation:





Plots on Imputation:





Brief Description of the Approach:

The overall strategy we employed to investigate both hypotheses includes conducting the test with and without considering missing values. We consider a confidence level of 95% for both the hypothesis, which means the α level is 0.05. Initially, we conduct hypothesis testing without addressing the missing data, meaning we exclude instances with missing values from the features and proceed with the test. Subsequently, addressing missing data involves managing two distinct types of missing values: MCAR, denoting missing values that are entirely random, and MNAR, signifying non-ignorable missing values where the missing data mechanism is connected to the absent values. For both types of missing values, we leverage the 'mice' R package to systematically generate and impute missing values. Following this imputation process, the hypothesis testing phase is performed.

Pearson Correlation

The Pearson correlation coefficient, often denoted by "r," is a measure of the linear relationship between two variables. The formula for calculating the Pearson correlation coefficient between two variables X and Y is as follows:

$$r = \sum(X_i - X_{mean})(Y_i - Y_{mean}) / \sqrt{}\sum(X_i - X_{mean})^2\sum(Y_i - Y_{mean})^2$$

- X_I are the X data points
- Y_i are the Y data points
- X_{mean} is the mean of the X data points
- Y_{mean} is the mean of the Y data points

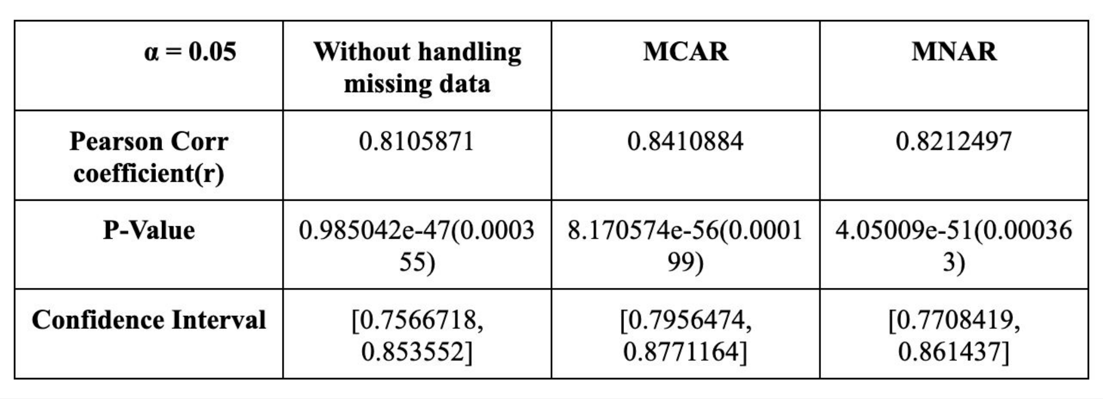
This formula calculates the correlation by dividing the covariance of the two variables by the product of their standard deviations. The resulting value ranges from -1 to 1, where: 1 means positive correlation, 0 means no correlation and -1 means negative correlation.

Hypothesis 1:

To examine the correlation between the horsepower and price of vehicles at a significance level(α) of 0.05

Null hypothesis(Ho): r=0

Alternate hypothesis(Ha) : r ≠ 0

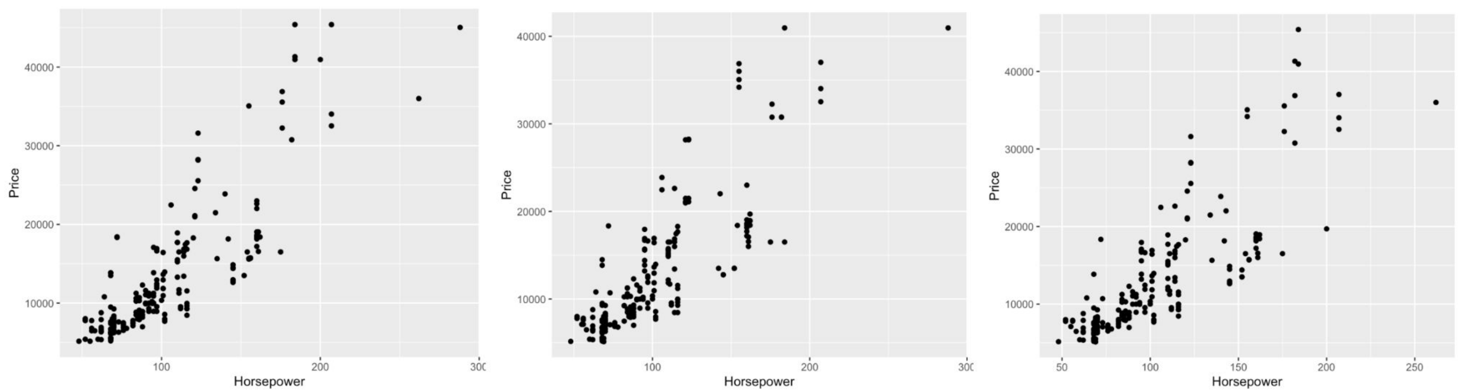


α = 0.05	Without handling missing data	MCAR	MNAR
Pearson Corr coefficient(r)	0.8105871	0.8410884	0.8212497
P-Value	0.985042e-47(0.000355)	8.170574e-56(0.000199)	4.05009e-51(0.000363)
Confidence Interval	[0.7566718, 0.853552]	[0.7956474, 0.8771164]	[0.7708419, 0.861437]

Examining the tabulated results provided, it is evident that in all cases, the P-value is below the significance level (α). This indicates that there is a significant correlation between horsepower and the price of the vehicle.

Plots for the first hypothesis

Below are the scatter plots corresponding to the three methods used for hypothesis testing.



MCAR

MNAR

Without handling missing data

Hypothesis 2:

We are interested in using the length, width, height, and wheelbase of a car to classify the number of doors of a car, which will be either 2 or 4. Under such scenarios, we build our logistic regression model by using the formula Y = number of doors which is a binary categorical variable, X_1 = length, X_2 = width, X_3 = height, and X_4 = wheelbase.

The second null hypothesis Ho: the coefficients of length, width, height, and wheelbase are equal to zero, indicating no effect of the corresponding predictor variable on the log-odds of the event. The alternative hypothesis is that the coefficient is not equal to zero, suggesting a significant effect. Here's a general outline of the hypothesis test:

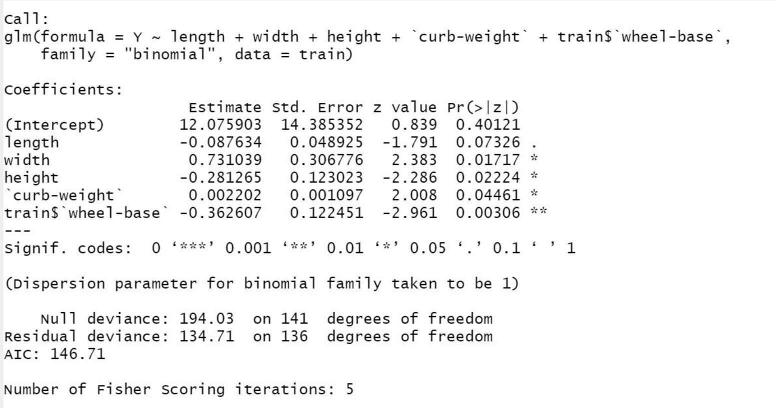
Ho: $\beta_1 = \beta_2 = \beta_3 = \beta_4$

Ha: at least one of B not equal to 0

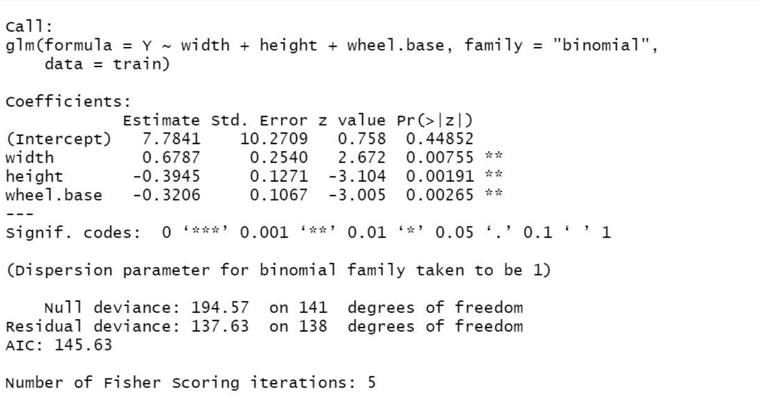
Approach:

1. Input original data and split into train and test subset
2. Build the logistic regression.
3. Call a stepwise function to identify the most relevant subset of predictor variables.
4. Compare two models using ANOVA
5. diagnose the logistic model, and check assumptions
6. Calculate the p-value and accuracy.
7. Get conclusion:

Logistic model before and after stepwise

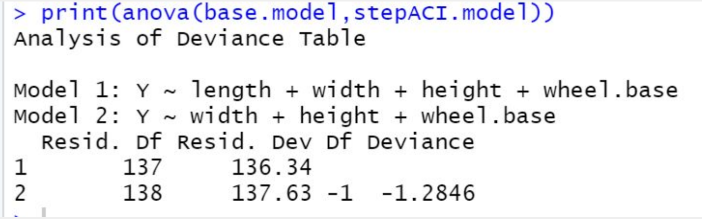


```
Call:
glm(formula = Y ~ length + width + height + `curb-weight` + train$`wheel-base`,
    family = "binomial", data = train)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        12.075903  14.385352   0.839  0.40121
length             -0.087634   0.048925  -1.791  0.07326 .
width               0.731039   0.306776   2.383  0.01717 *
height             -0.281265   0.123023  -2.286  0.02224 *
`curb-weight`       0.002202   0.001097   2.008  0.04461 *
train$`wheel-base` -0.362607   0.122451  -2.961  0.00306 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 194.03  on 141  degrees of freedom
Residual deviance: 134.71  on 136  degrees of freedom
AIC: 146.71

Number of Fisher Scoring iterations: 5
```



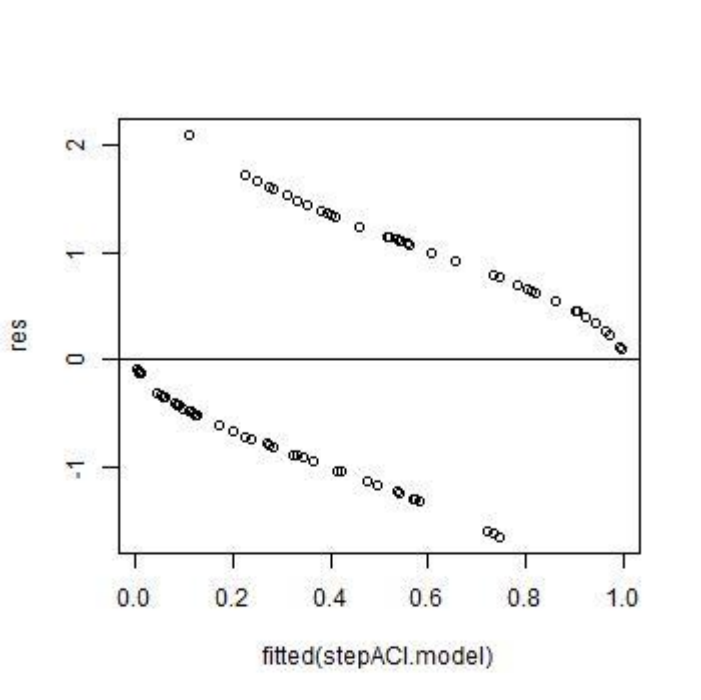
```
Call:
glm(formula = Y ~ width + height + wheel.base, family = "binomial",
    data = train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.7841    10.2709   0.758  0.44852
width         0.6787     0.2540   2.672  0.00755 **
height       -0.3945     0.1271  -3.104  0.00191 **
wheel.base   -0.3206     0.1067  -3.005  0.00265 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 194.57  on 141  degrees of freedom
Residual deviance: 137.63  on 138  degrees of freedom
AIC: 145.63

Number of Fisher Scoring iterations: 5
```

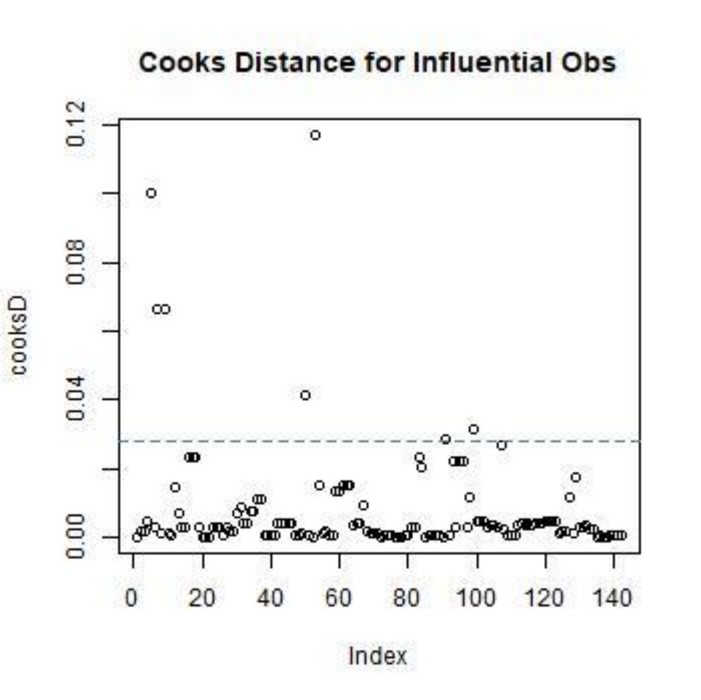
ANOVA:

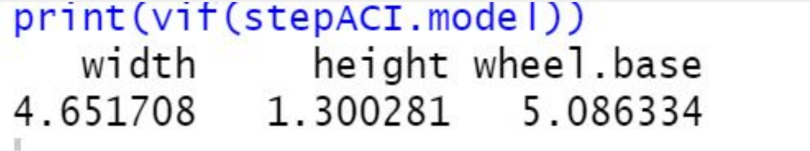


```
> print(anova(base.model,stepACI.model))
Analysis of Deviance Table

Model 1: Y ~ length + width + height + wheel.base
Model 2: Y ~ width + height + wheel.base
  Resid. Df Resid. Dev Df Deviance
1       137     136.34
2       138     137.63 -1  -1.2846
```

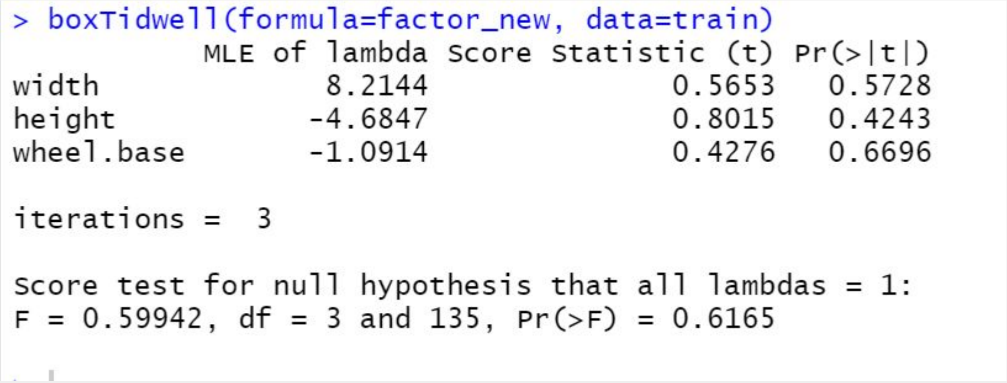
Diagnosis of the model





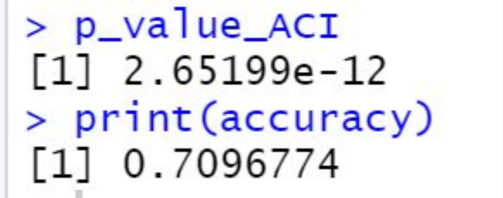


```
print(vif(stepACI.model))
   width     height wheel.base 
4.651708   1.300281   5.086334 
```



```
> boxTidwell(formula=factor_new, data=train)
           MLE of lambda Score Statistic (t) Pr(>|t|)
width             8.2144              0.5653   0.5728
height           -4.6847              0.8015   0.4243
wheel.base       -1.0914              0.4276   0.6696

iterations =  3 

Score test for null hypothesis that all lambdas = 1:
F = 0.59942, df = 3 and 135, Pr(>F) = 0.6165
```

P-Value and Accuracy



```
> p_value_ACI
[1] 2.65199e-12
> print(accuracy)
[1] 0.7096774
```

P value is less than type I error 0.05, so we reject the null hypothesis and conclude that variable width, height, and wheelbase have significant effects. The variable length has no effect because it is dropped by a stepwise approach.

MCAR vs MNAR vs ORIGINAL

	MCAR	MNAR	ORIGINAL
P-value	4.181877e-12	2.126188e-12	2.65199e-12

	MCAR	MNAR	ORIGINAL
Accuracy	0.6935484	0.6935484	0.7096774

Conclusion

Concluded from the plots and results, the MCAR data had the highest ACI, followed by the Original missing value data and the MNAR data. Both MCAR and MNAR data increase the collinearity between variables, as you can see, the wheel. base value increases from 5.08 to 5.14 and 5.18. The MCAR data has the biggest p-value, followed by MNAR and Original, but all of them are smaller than type I error, 0.05, so we reject the null hypothesis for both the hypothesis and conclude that there is a significant correlation between horsepower and price and then variable width, variable height, and variable wheelbase have significant effect. Finally, MCAR and MNAR data also happen to decrease the accuracy in the case of the second hypothesis.

Thank You