Introduction:

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 field of car attributes.

The first hypothesis examines the potential correlation between a vehicle's horsepower and its price. Horsepower, a pivotal metric reflecting an engine's power output, is often considered a primary factor influencing a vehicle's overall performance and desirability. The purpose of this study is to determine whether there is a statistically significant relationship between car pricing and horsepower.

The second hypothesis investigates the difference in mean fuel efficiency between gas-powered and diesel-powered vehicles. With an increasing emphasis on environmental sustainability and fuel economy, understanding the differences in efficiency across these fuel types is imperative. Through hypothesis testing, we aim to determine if there is a statistically significant divergence in the mean fuel efficiency of gas and diesel vehicles.

In the subsequent sections of this project, we will delve into the methodologies employed for data collection, exploratory data analysis, interpretation of results, and conclusions. The findings from these hypotheses can be a valuable insight into the automobile industry,

Data Description:

The dataset utilized in this project has been sourced from the UC Irvine Machine Learning Repository and comprises 204 entries with 25 distinct features. These features encompass a comprehensive range of automobile information, including car make, fuel type, dimensions (height and weight), engine size, horsepower, vehicle price, and miles per gallon (mpg). The dataset thus provides a multifaceted perspective on various attributes associated with the studied cars, forming the foundation for our exploration. The description of some important features is as follows:

Height: The height of the car.
Weight: The weight of the car.
Horsepower: Indicates how quickly the force is produced from a vehicle's engine.
Compression ratio: The ratio between the volume of the cylinder with the piston in the bottom position and the top position.
Wheelbase: The distance between the centers of the front and rear wheels.
Stroke: The distance traveled by the piston during each cycle.
Price: The price of the vehicle.
Symboling: Corresponds to a car’s insurance risk level.
Number of doors: The number of doors in the car
Fuel type: The fuel type of the vehicle such as gas or diesel

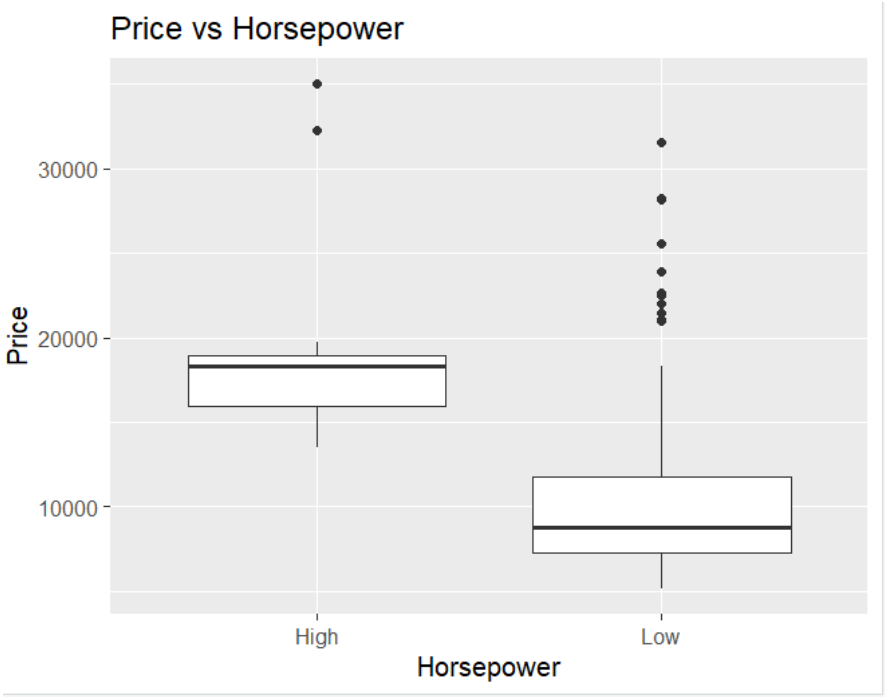
City mpg: Miles the vehicle can travel inside the city per gallon.
highway mpg: Miles the vehicle can travel on the highway per gallon.
Normalized loss: The relative average loss payment per insured vehicle.

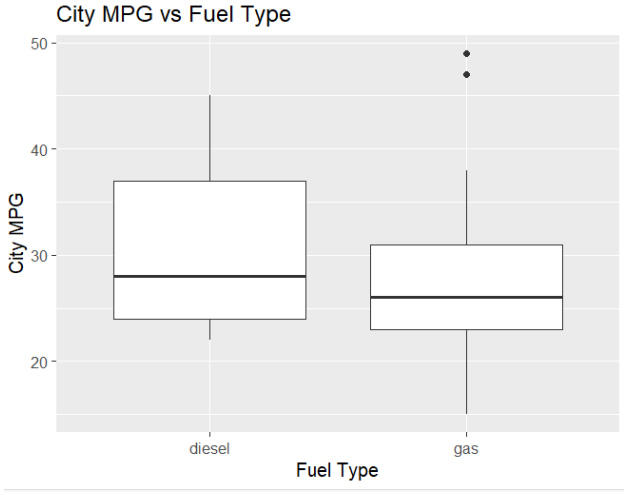
Brief Description of the Approach:

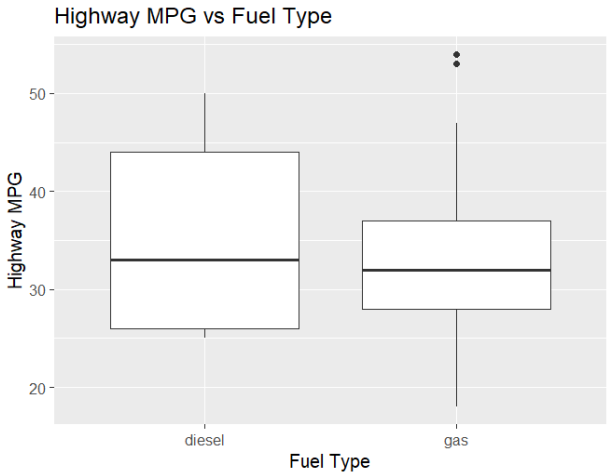
The overall strategy we employed to investigate both hypotheses includes conducting the test with and without considering missing values. We consider a confidence level of 95% for both the hypothesis, which means the α level is 0.05. Initially, we conduct hypothesis testing without addressing the missing 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.

Exploratory Data Analysis:

All the exploratory data analysis is done on the data set by removing all the rows that had at least 1 missing value. Below are box plots for price vs horsepower, City mpg vs Fuel type, and Highway mpg vs fuel type. For the context of analysis, the horsepower values are grouped into two sections. Values of more than 150 were put in high and the rest into low.







Based on the box plots provided, it is evident that there is a noticeable trend between Price and horsepower. The median, 25th percentile, and 75th percentile prices for vehicles with high horsepower compared to those with low horsepower demonstrate a noticeable gap. In contrast, the remaining plots lack a similar level of significance in the observed trend. Consequently, we sought to formulate our hypothesis for the test by examining the correlation between the horsepower and the price of the vehicle.

Missing Values:

To investigate the effect of the missing value, we create missing values in types of MCAR and MNAR via the "*ampute*" function from the MICE package for both hypotheses. Twenty percent of the data will be missed after we apply the function. After a scale of the dataset is missing, missing data will be imputed by the *"mice"* (multivariate imputation by chained equation) function from MICE. The imputation method will be set with a method of predictive mean matching. This method will get the correlation between the columns within the hypothesis and predict the missing value based on it. Then, both hypotheses will be analyzed with the original dataset, a dataset imputed after MCAR, and a dataset imputed after MNAR.

Note: Since there are some bugs from the "*ampute*" function to generate missing values in the MNAR method, which is hard for us to handle, it can't generate missing values in the MNAR method if the original dataset has already exist missing values. So, before generating the missing value, the columns that will be analyzed in the hypothesis were updated with a linear regression model to fill up the original missing value to make "*ampute*" work. And just for clarification, the original dataset in both hypotheses refers to the original dataset which used

list-wised deletion to ignore the missing value in the original dataset and without other manipulation.

Methods used for Hypothesis:

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:

$$r = \sum(Xi - Xmean)(Yi - Ymean) / \sqrt{\sum(Xi - Xmean)2\sum(Yi - Ymean)2}$$

- $Xi\ are\ the\ X\ data\ points$
- $Yi\ are\ the\ Y\ data\ points$
- $Xmean\ is\ the\ mean\ of\ the\ X\ data\ points$
- $Ymean\ is\ the\ mean\ of\ the\ Y\ data\ points$

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.

Logistic regression:

Logistic regression is a statistical method used for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is typically a binary variable. It is a type of regression analysis that is appropriate when the dependent variable is categorical.

The logistic regression model is a linear combination of variables, but it applies a logistic function to this combination:

$$\text{logit}(p) = \log(p / 1 - p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Where:
- p is the probability of the dependent variable equaling a case
- $\beta_0, \beta_1, ..., \beta_n$ are the coefficients.
- $x_1, x_2, \ldots, x_n$ are the independent variables.

Interpretation:
- The coefficients (beta values) indicate the change in the log odds of the outcome for a one-unit increase in the predictor variable.

Assumptions:
1. The observations are independent.
2. The model should have little or no multicollinearity.
3. Logistic regression requires large sample sizes.
4. There are no extreme outliers.
5. There is a linear relationship between the predictor variables and the logit of the response variable.

Stepwise Logistic regression:

Stepwise logistic regression is a statistical technique used for building a logistic regression model by iteratively selecting and excluding predictor variables based on their statistical significance. The purpose of stepwise regression is to identify the most relevant subset of predictor variables that contribute significantly to the prediction of the binary outcome variable.

The stepwise logistic regression process typically involves two main steps: forward selection and backward elimination.

Forward Selection:

- Start with an empty model and evaluate each predictor variable individually based on a predetermined criterion (e.g., p-value, likelihood ratio test, Akaike Information Criterion).
- Select the variable with the lowest criterion value and add it to the model.
- Repeat this process iteratively, adding one variable at a time, until no more variables meet the inclusion criteria.

Backward Elimination:

- Start with the full model containing all predictor variables.
- Evaluate each variable's contribution to the model based on the chosen criterion.
- Remove the variable with the highest criterion value if it exceeds a predetermined significance level.
- Iteratively eliminate variables until no more variables meet the exclusion criteria.

The stepwise process continues alternating between forward selection and backward elimination until no more variables can be added or removed based on the specified criteria.

First Hypothesis:

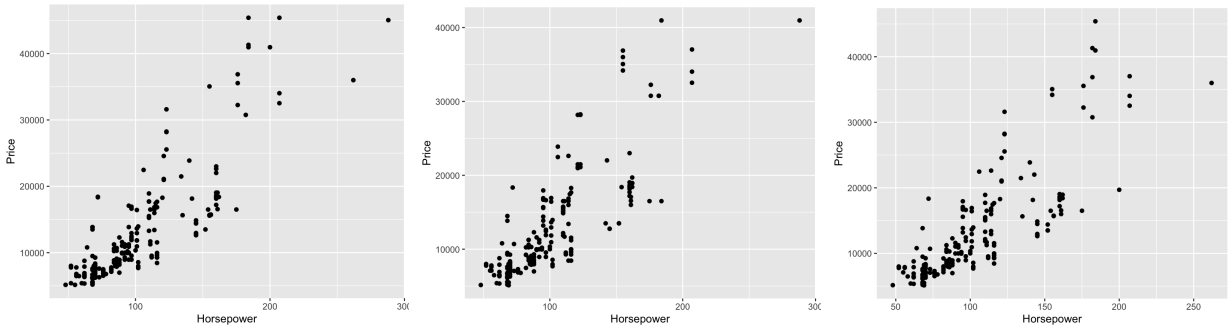
As we mentioned earlier, the first hypothesis aims to examine the correlation between the horsepower and price of vehicles. Across all three scenarios—without any handling, under the MCAR type, and within the MNAR type—we used the Pearson correlation metric with a confidence level of 95%. Which means the α level is 0.05. The null hypothesis states that there is no significant correlation between horsepower and price while the alternative hypothesis states there is a significant correlation.

$$Null\ hypothesis(Ho) : r = 0$$
$$Alternate\ hypothesis(Ha) : r \neq 0$$
$$Significance\ level(\alpha) = 0.05$$

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. It is also evident that the Pearson correlation coefficient is highest for the MCAR type, followed by the MNAR type and the scenario without missing values. Additionally, the endpoints of the confidence interval for the MCAR type are higher compared to the other two cases.

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



MCAR MNAR Without handling missing data

Second Hypothesis:

In logistic regression, hypothesis tests are used to assess the statistical significance of the coefficients associated with the predictor variables. The logistic regression model estimates the probability of an event occurring, and the coefficients represent the change in the log odds of the event for a one-unit change in the corresponding predictor variable.

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 is 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:

$$Ho: \ \beta 1 = \beta 2 = \beta 3 = \beta 4 = 0$$

$$Ha: \ at\ least\ one\ of\ \beta\ not\ equal\ to\ zero.$$

The test statistic is calculated based on the estimated coefficient, its standard error, and the assumed distribution. Wald test is often used for hypothesis testing in logistic regression. Since the Wald test is asymptotically distributed as a chi-squared distribution with one degree of freedom under the null hypothesis, we use Chi-square statistics to calculate the p-value and compare it with our type 1 error. If the p-value is bigger than the type 1 error, there is not enough evidence to reject Ho, and otherwise, reject Ho when the p-value is smaller than the type 1 error.

The whole process of the test approach is shown as follows (for original data):

1. Input original data and split into train and test subset
2. Build the logistic regression. Here is the summary of our models:

```
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
```

3. Call a stepwise function to identify the most relevant subset of predictor variables. Here is the result. The length variable is drop, which means variable length does not contribute to the prediction of the binary outcome variable.

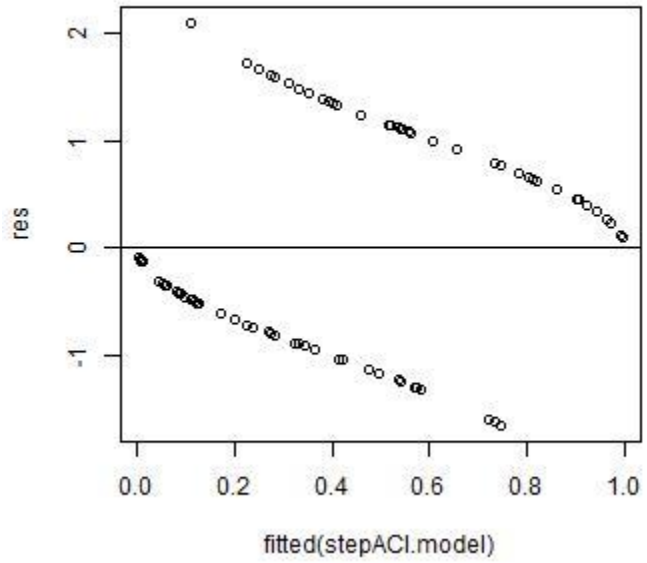
```
Call:
glm(formula = Y ~ width + height + wheel.base, family = "binomial",
    data = train)

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
```

4. Compare two models using ANOVA. The resid is close, 136.34 compared with 137.63, so variable length is useless.

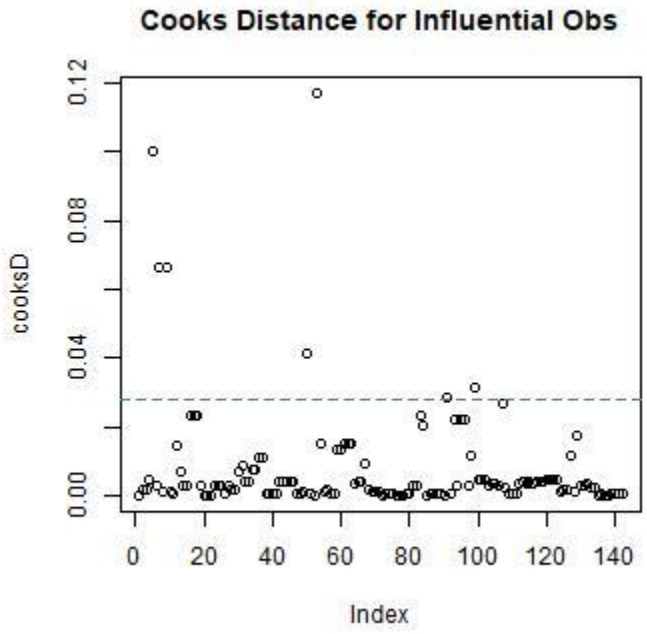
```
> print(anova(base.model,stepACI.model))
Analysis of Deviance Table

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
```

5. diagnose the logistic model, and check assumptions:

The residual plot shows that observations are approximately independent.



The cookie distance shows that there are a few outliers.



The collinearity test shows that there are moderate correlations between variables because the wheelbase is 0.08 bigger than 5.

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

The p-value of width, height, and wheelbase is bigger than 0.05, so there is a linear relationship between the logit(P) and the number of doors.

```
> 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
```

6. Calculate the p-value and accuracy. Accuracy is calculated by using test data:

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

7. Get conclusion:
   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.

Then we do the same process for the MCAR and MNAR data and compare it with the original data:
Mcar vs Mnar vs Original data
stepACI model of Mcar:

```
[1] "summary stepACI model"

Call:
glm(formula = factor_new, family = "binomial", data = train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.3433    10.1338   0.626  0.53134
width         0.6969     0.2588   2.692  0.00710 **
height       -0.3752     0.1211  -3.098  0.00195 **
wheel.base   -0.3289     0.1088  -3.023  0.00250 **
---
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: 138.56  on 138  degrees of freedom
AIC: 146.56

Number of Fisher Scoring iterations: 5
```

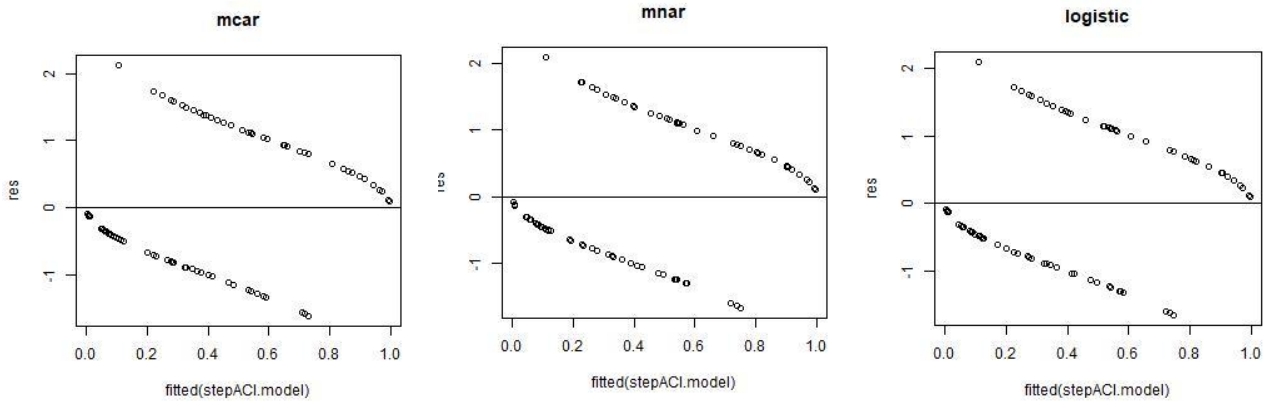
stepACI model of Mnar:

```
[1] "summary stepACI model"

Call:
glm(formula = factor_new, family = "binomial", data = train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   9.2186    10.2022   0.904  0.36621
width         0.6551     0.2525   2.594  0.00948 **
height       -0.4005     0.1260  -3.178  0.00148 **
wheel.base   -0.3163     0.1076  -2.939  0.00330 **
---
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: 137.18  on 138  degrees of freedom
AIC: 145.18

Number of Fisher Scoring iterations: 5
```

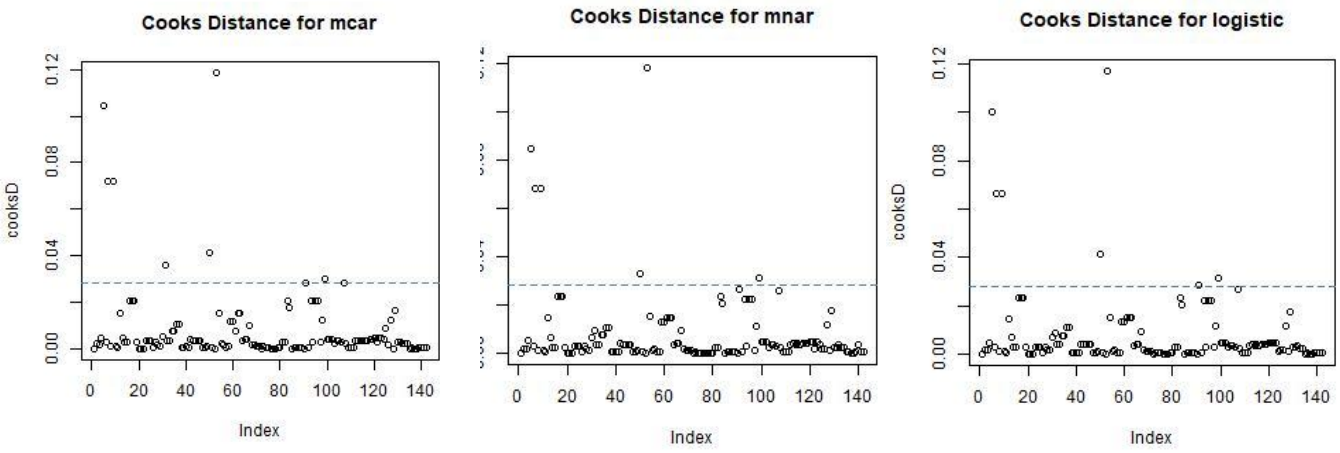
stepACI model of Origin:

```
[1] "summary stepACI model"

Call:
glm(formula = factor_new, 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
```

Independence test:



Cook's distance test:



collinearity assumption test

	width	height	wheel.base
MCAR	4.794961	1.291382	5.187811
MNAR	4.696617	1.284548	5.141819
ORIGINAL	4.651708	1.300281	5.086334

box-Tidwell test:

	Width Pr(>|t|)	Height Pr(>|t|)	Wheel.base Pr(>|t|)
MCAR	0.6649	0.6175	0.5482
MNAR	0.6342	0.4459	0.6091
ORIGINAL	0.5728	0.4243	0.6696

P-value:

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

Accuracy of test data:

	MCAR	MNAR	ORIGINAL
Accuracy	**0.6935484**	**0.6935484**	**0.7096774**

Conclusion:

Concluded from the plots and table, 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 in the table collinearity assumption test, the wheel. base value increases from 5.08 to 5.14 and 5.18. The MCAR data has the biggest p-value, followed by Original and MNAR, but all of them are smaller than type I error, 0.05, so we reject the null hypothesis for both two hypothesis and conclude that there is a significant correlation between horsepower and price and then variable width, variable height, and variable wheelbase have a significant effect when classifying number of doors. Finally, MCAR and MNAR data also happen to decrease the accuracy in the case of the second hypothesis.

Appendix

Original dataset link: Automobile - UCI Machine Learning Repository

References for computing the Missing values:

1. Missing-data imputation
2. Tutorial on 5 Powerful R Packages used for imputing missing values

h1_mcar.csv data for hypothesis 1 with MCAR.
h1_mcar_imputed.csv data imputed with MCAR values for hypothesis 1
h1_mnar.csv data for hypothesis 1 with MNAR.
h1_mnar_imputed.csv data imputed with MNAR values for hypothesis 1
h2_mcar.csv data for hypothesis 2 with MCAR.
h2_mcar_imputed.csv data imputed with MCAR values for hypothesis 2
h2_mnar.csv data for hypothesis 2 with MNAR.
h2_mnar_imputed.csv data imputed with MNAR values for hypothesis 1