MH3511 Data Analysis with Computer

Group Project

Analysing Pricing Dynamics and other Key Attributes of Ford Vehicles

Name	Matriculation Number
Choo Yi Ken	U2240710B
Matthew Heng Yu Jie	U2223483D
Tong Hao Kit	U2240130E
Grand Tan Ze Ming	U2240872B
Hydee Qurniawan B Rosli	U2040911F

Abstract:

With the surge of demand for vehicles from around the world today, the sales from Ford cars have been at its highest. As such, this report aims to provide a thorough examination of Ford car pricing based on different key attributes by using basic data analysis techniques. Results show moderate correlations between the prices of cars and characteristics such as mpg and mileage. The findings offer detailed insights into Ford car pricing trends and their changes as time goes on. Suggestions for further research and real-world implementations of the results are also addressed.

Table of contents

1. Introduction

Ford Motor Company is a globally recognised US-based vehicle manufacturer, with almost 2 million units being sold in 2023 alone (Wayland, 2024). Profits from Ford vehicle sales have generally risen over the years, with the company reportedly generating a revenue of nearly 176.2 billion U.S. dollars in 2023 (Statista,2024). This may be due to the rising demand for Ford F-series trucks (Tech Xplore, 2023) and crossover SUVs (The Business Times, 2024) over the years.

As such, this project aims to study the trend of prices of used Ford cars over a time period and to determine any associations between its price and other attributes. Subsequently, a large dataset containing the attributes of used Ford cars recorded over multiple years was utilised.

Specifically, the project seeks to address the following questions:

1. Is the price of a Ford car dependent on any of its attributes (model, transmission type, mileage, fuel type or fuel consumption rate)?
2. Does time (year) affect Ford car prices?
3. Is there an attribute of Ford cars that influences the distribution of prices to a greater extent than the other attributes? By how much?

The dataset will be analysed statistically using R language, and so this report will project the findings in the next few sections, along with explanations and relevant conclusions.

2. Data Description

The dataset, named “ford.csv” was derived from Kaggle.com, which is an online data repository for machine learning. The dataset is made up of only one data frame consisting of data from used Ford vehicles recorded over the years between 1996 and 2020. Each row represents a unit of used Ford car and its corresponding attributes.

Before further analysis, some initial adjustments were made to the dataset:

- Removed one row of which the value under the column “year” is “2060”, as it was observed to be a lone anomaly of the dataset.
- Columns “tax” and “engineSize” were removed as they were deemed irrelevant for this project.

Altogether, there were a total of 17965 observations (Ford cars) with their attributes being split into 7 variables of different types:

1. *model* : Model of car
2. *year* : Year of which the car was manufactured
3. *price* : resale price of car at the point of observation

4. *transmission* : Type of transmission (gearbox) that is inbuilt in car
5. *mileage* : Total distance covered by car in its lifetime till the point of observation
6. *fuelType* : Type of fuel required for car to operate
7. *mpg* : stands for "Miles per Gallon", that is, how far the car can travel with 1 gallon of fuel (measures car's fuel consumption rate)

The next section will address these variables in greater detail, with further necessary adjustments to the data being made before performing proper statistical tests.

3. Description and Cleaning of Dataset

In this section, preliminary investigations of the following variables in the dataset were performed, namely:

1) Main variable of interest: *price*
2) Other variables: *model*, *transmission, fuelType, mileage, mpg, year*

Here, the main aim is to remove anomalies from the dataset (if there are any), and to check for skewness and make necessary data transformations where needed.

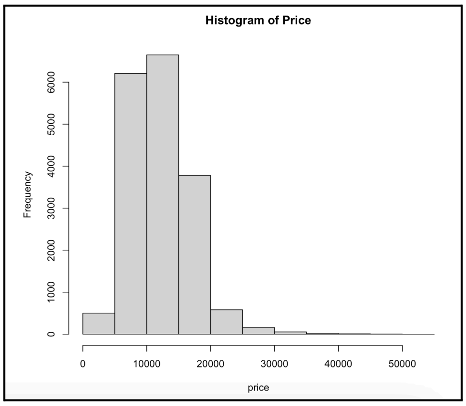
3.1 Summary statistics for the main variable of interest

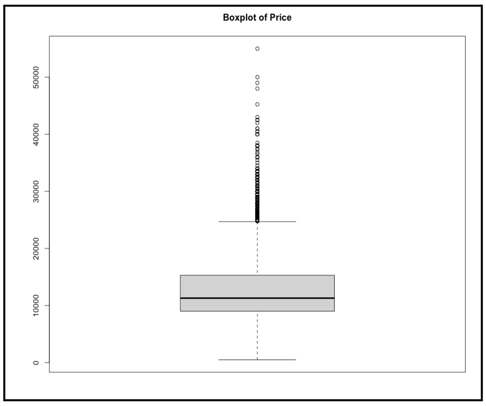
3.1.1 Main Variable of Interest : price

The summary of *price* is shown below :

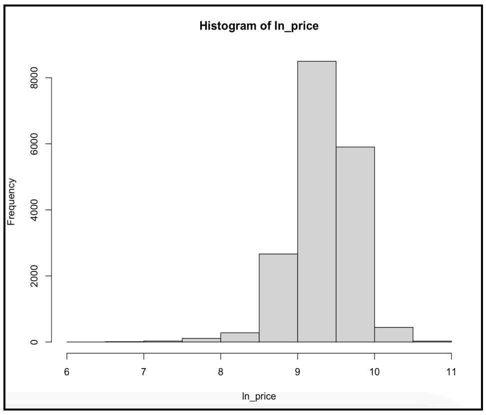
```
> summary(car_data$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    495    8999   11291   12280   15299   54995 
```

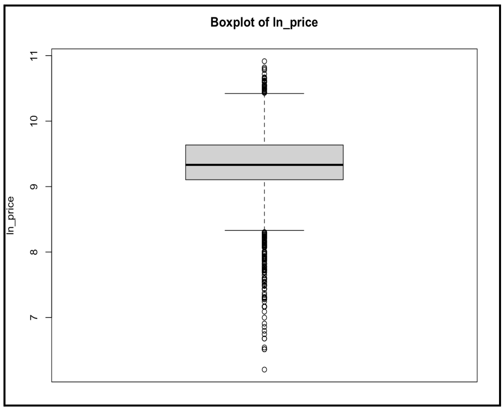
The histogram below shows the distribution for the variable *price*.



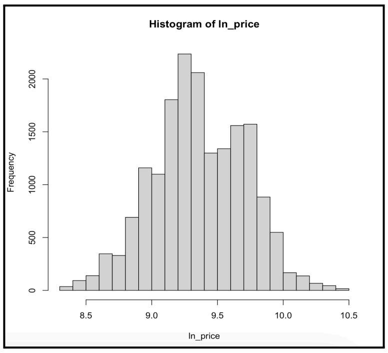


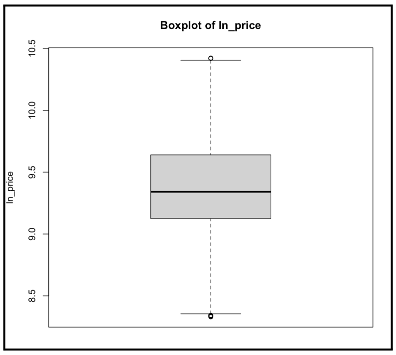
As shown in the figure above, the variable *price* is highly skewed to the right. Therefore, adjustments were made to the dataset to reduce its skewness. A ln-transformation is performed on the dataset and the resulting histogram and boxplot of *ln_price* are plotted as shown below.

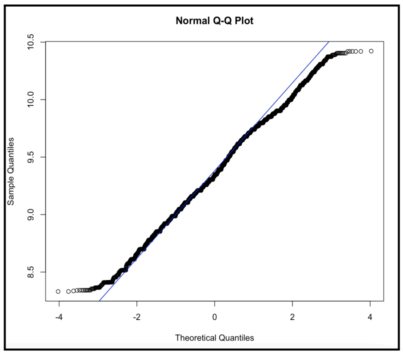




It is observed that the *ln_price* dataset is now more symmetric but the ln-transformed data appears to have some outlying values at the left tail. Thus, the anomalies of the dataset *ln_price* are removed using the boxplot rule for outliers. As a result, there is approximately 1.89% of the data being removed. The resulting histogram, boxplot, and qq plot of the filtered dataset are shown below. We shall proceed to the next section with this filtered dataset.







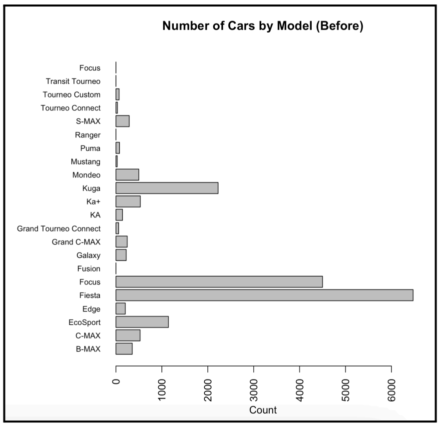
3.2 Summary statistics for the other variables

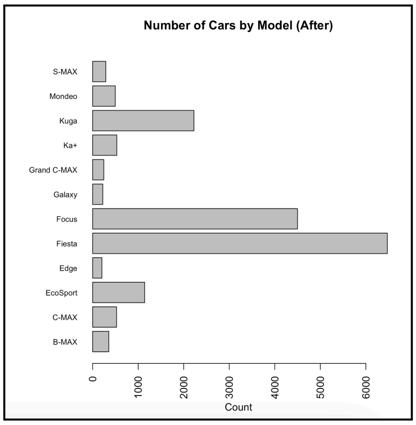
In this section, the 6 other variables : *model*, *transmission, fuelType, mileage, mpg, year* are examined and necessary changes were implemented to the data stored in these variables to ensure minimal skewness and removal of anomalies.
For each subsection, the histogram , boxplot , applied transformations and anomalies removed (if applicable) are shown.

3.2.1 Variable 1 : model

The *model* variable represents the model of a Ford car. Below is the barplot of the variable *model*. It can be observed that the mode of this dataset is the model “Fiesta”. Since there is a huge difference in the number of cars among different models, thus, cars with a model consisting of less than 200 cars in the dataset are omitted. A total of 199 (1.13%) cars were omitted as a result.

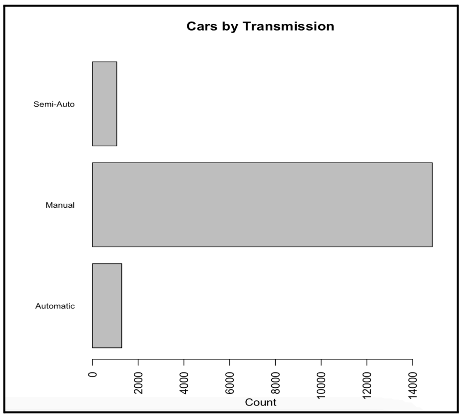
.





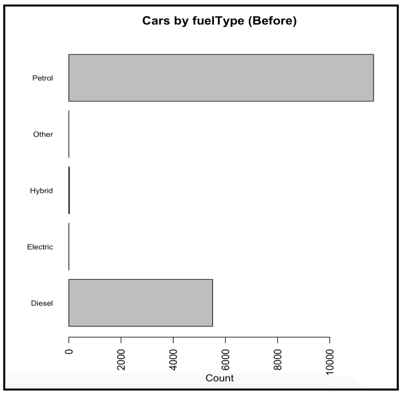
3.2.2 Variable 2 : Transmission

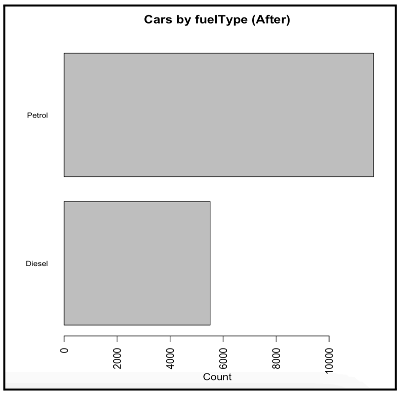
The *transmission* variable represents the type of transmission of a car. No changes (In transformation/removal of anomalies/trimming of dataset) were made. Below is the barplot of the variable *transmission*. It can be observed that the mode of this dataset is the “Manual” transmission.



3.2.3 Variable 3 : fuelType

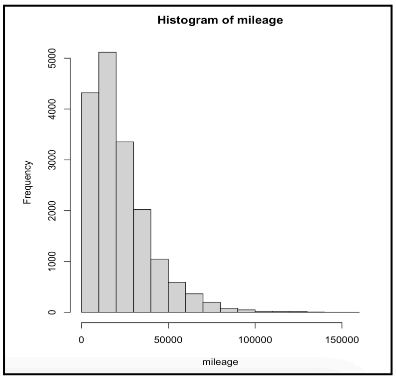
The *fuelType* variable represents the type of fuel used by each car. Since there is a huge difference in the number of cars among different fuel types, the fuel types of “electric”, “hybrid”, and “other” are removed. As a result, there are 24 (0.14%) of cars omitted.

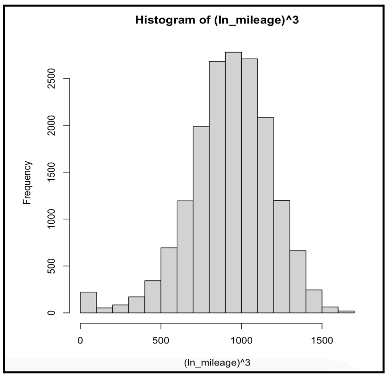


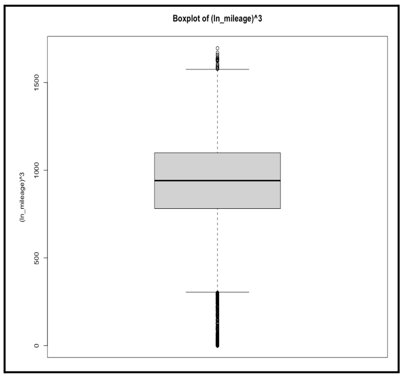


3.2.4 Variable 4 : mileage

The *mileage* variable represents the size of the mileage of a car. The *mileage* data is right skewed, therefore a ln-transformation is applied followed by a cube transformation to make it more symmetry and normally distributed.

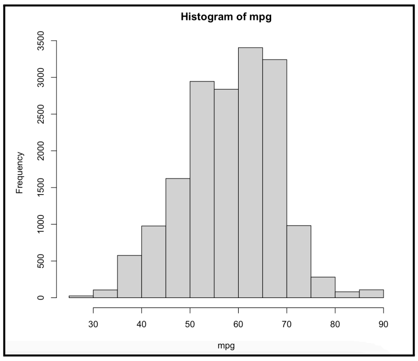


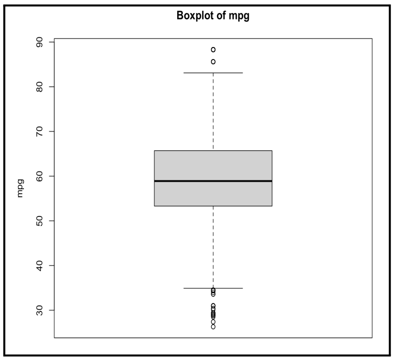




3.2.5 Variable 5 : mpg

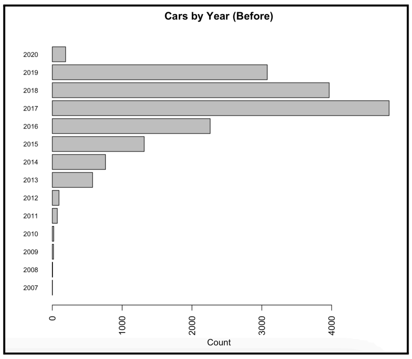
The *mpg* variable is the mile per gallon a car is able to travel. No changes (ln transformation/removal of anomalies/trimming of dataset) were made.

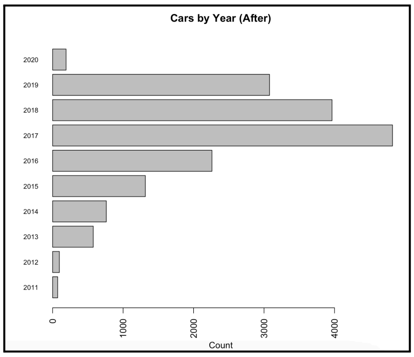




3.2.6 Variable 6 : year

The *year* variable represents the year a car is manufactured. Since there is a huge difference in the number of cars among different years, cars manufactured in the years 2007 to 2010, each having fewer than 30 cars, are eliminated as their number of cars were too small compared to other years. As a result, 47 (0.27%) of cars were omitted.





3.3 Final Dataset for Analysis

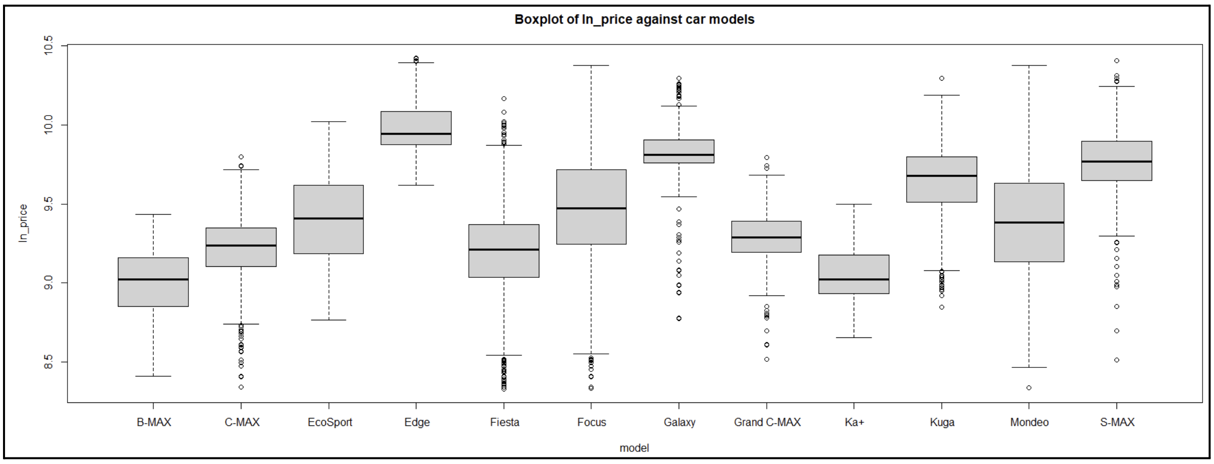
The variables that required necessary changes before analysis are *price, transmission, fuelType, mileage, model, and year*. There are no changes applied for the *mpg* variable.
Overall, the dataset has reduced in size by 4.46 % due to the removal of outliers, and is now more symmetric. Therefore, we can now proceed to use the aforementioned dataset for analysis.

4. Statistical Analysis

4.1 Statistical Tests

4.1.1 Relation between ln_price and model

For section 4.1.1, we want to determine whether the resale price of a Ford car depends on its model. Since *model* is a categorical variable with 12 categories, we use an analysis of variance (1-way ANOVA) test to determine whether *ln_price* has much difference between different car models. The plot below illustrates the distribution of *ln_price* for different car models.



By visualisation, the spread of *ln_price* is not quite the same for all 12 car models.
By hypothesis testing (ANOVA test):

H_0: $\mu_{B\text{-}MAX} = \mu_{C\text{-}MAX} = \mu_{EcoSport} = \mu_{Edge} = \mu_{Fiesta} = \mu_{Focus} = \mu_{Galaxy} = \mu_{Grand\ C\text{-}MAX} = \mu_{Ka+} = \mu_{Kuga} = \mu_{Mondeo} = \mu_{S\text{-}MAX}$

H_1: not all μ_i are equal

```
                             Df Sum Sq Mean Sq F value Pr(>F)    
factor(car_data$model)       11  649.3   59.02     757 <2e-16 ***
Residuals                 17129 1335.5    0.08                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test returns a p-value of <2e-16, which is less than 0.05 at significance level of 0.05. Thus, H_0 is rejected, we conclude that resale price of a Ford branded car is dependent on the car model.

Since the null hypothesis is rejected, we use pairwise.t.test() to determine which car model group is different from the other groups.

```
             Pairwise comparisons using t tests with pooled SD 

data:  car_data$ln_price and car_data$model 

             B-MAX   C-MAX   EcoSport  Edge    Fiesta  Focus   Galaxy  Grand C-MAX  Ka+     Kuga    Mondeo 
 C-MAX       < 2e-16 -       -         -       -       -       -       -            -       -       -      
 EcoSport    < 2e-16 < 2e-16 -         -       -       -       -       -            -       -       -      
 Edge        < 2e-16 < 2e-16 < 2e-16   -       -       -       -       -            -       -       -      
 Fiesta      < 2e-16 0.77239 < 2e-16   < 2e-16 -       -       -       -            -       -       -      
 Focus       < 2e-16 < 2e-16 2.2e-07   < 2e-16 < 2e-16 -       -       -            -       -       -      
 Galaxy      < 2e-16 < 2e-16 < 2e-16   2.4e-14 < 2e-16 < 2e-16 -       -            -       -       -      
 Grand C-MAX < 2e-16 0.00036 4.7e-10   < 2e-16 5.3e-05 < 2e-16 < 2e-16 -            -       -       -      
 Ka+         0.00210 < 2e-16 < 2e-16   < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16      -       -       -      
 Kuga        < 2e-16 < 2e-16 < 2e-16   < 2e-16 < 2e-16 < 2e-16 1.0e-14 < 2e-16      < 2e-16 -       -      
 Mondeo      < 2e-16 < 2e-16 0.28884   < 2e-16 < 2e-16 1.9e-06 < 2e-16 1.3e-06      < 2e-16 < 2e-16 -      
 S-MAX       < 2e-16 < 2e-16 < 2e-16   < 2e-16 < 2e-16 < 2e-16 0.11690 < 2e-16      < 2e-16 7.9e-11 < 2e-16

P value adjustment method: none 
```

By the above result, the null hypothesis is rejected for most of the car model pairs at significance level 0.05. However there are 3 pairs that have strong evidence supporting that they have equal mean, which are:

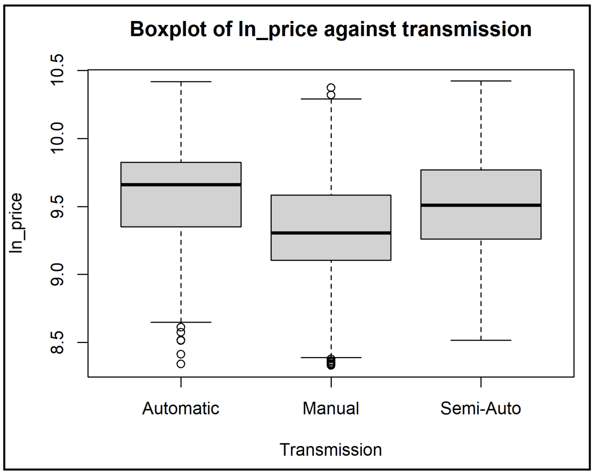
i) C-MAX and Fiesta with a p-value of 0.7724 (>0.05)

ii) Mondeo and EcoSport with a p-value of 0.2888 (>0.05)

iii) S-MAX and Galaxy with a p-value of 0.1169 (>0.05)

4.1.2 Relation between ln_price and transmission

For section 4.1.2, we want to determine whether the resale prices of Ford cars depend on the transmission of the car. We performed an analysis of variance (1-way ANOVA) test to determine whether *ln_price* has the same mean values for all 3 types of transmission method. The plot below illustrates the distribution of *ln_price* of 3 different transmissions.



By visualisation,the spread of *ln_price* is not the same for all 3 types of transmissions. It is obvious that the mean car_price of Manual is the smallest, followed by Semi-Auto and Automatic.

By hypothesis testing (ANOVA test):

H_0: $\mu_{auto} = \mu_{manual} = \mu_{semi\text{-}auto}$

H_1: not all μ_i are equal

```
                               Df Sum Sq Mean Sq F value Pr(>F)
factor(car_data$transmission)     2  108.7   54.35   496.5 <2e-16 ***
Residuals                     17138 1876.1    0.11
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test returns a p-value of <2e-16, which is less than 0.05 at significance level of 0.05. Thus, H_0 is rejected, we conclude that the prices of cars are dependent on the car's transmission method.

Since the null hypothesis is rejected, we use pairwise.t.test() to determine which transmissions have the same means.

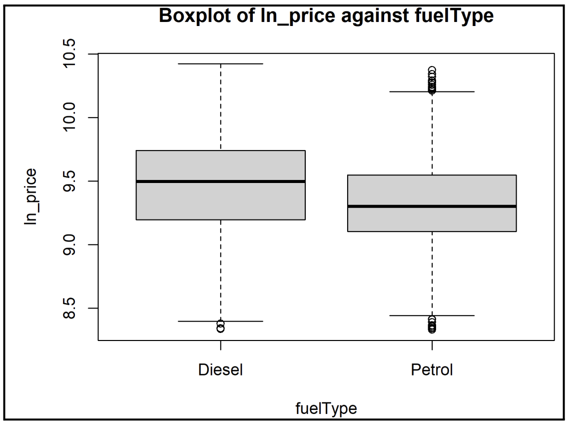
```
       Pairwise comparisons using t tests with pooled SD

data:  car_data$ln_price and car_data$transmission

          Automatic Manual
Manual    < 2e-16   -
Semi-Auto 4.7e-08   < 2e-16

P value adjustment method: none
```

By the above result, the null hypothesis is rejected for all the car transmission pairs at significance level 0.05. Hence, we conclude that $\mu_{Auto} \neq \mu_{Manual} \neq \mu_{Semi\text{-}auto}$.

4.1.3 Relation between ln_price and fuel_type

For section 4.1.3, we want to determine whether resale prices of cars depend on the type of fuel that the car requires to run. For simplicity purposes, we removed categories (within *fuelType*) such as ‘other’, ‘hybrid’ and ‘electric’, and we were consequently left with two dominant fuel types: ‘petrol’ and ‘diesel’.

Firstly, we performed F-test to check if we could assume the variance of the two categories are the same. Next, we performed t.test() to find out if the two categories have the same mean. The plot below illustrates the distribution of *ln_price* of cars that run on two different fuel types.



By visualisation, the mean values for the *ln_price* of different types of fuel are not quite the same.

By hypothesis testing (F-test):

H_0: $\sigma^2_{Diesel} = \sigma^2_{Petrol}$

H_1: $\sigma^2_{Diesel} \neq \sigma^2_{Petrol}$

```
>var.test(car_data[car_data$fuelType=="Diesel",10],car_data[car_data$fuelTy
pe == "Petrol",10])

F test to compare two variances

data:         car_data[car_data$fuelType     ==     "Diesel",     10]     and
car_data[car_data$fuelType == "Petrol", 10]
F = 1.3503, num df = 5489, denom df = 11650, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.290642 1.413267
sample estimates:
ratio of variances
           1.35031
```

By the above result, F-test returns a p-value of 2.2e^-16, which is less than 0.05 at significance level of 0.05. Thus, H_0 is rejected and we cannot assume that the variances are the same for the samples with fuelType “diesel” and “petrol”.

Next, we proceed with t-test (assuming that variances are not equal for both samples)

By t-test, hypothesis testing:

H_0: $\mu_{Diesel} = \mu_{Petrol}$

H_1: $\mu_{Diesel} \neq \mu_{Petrol}$

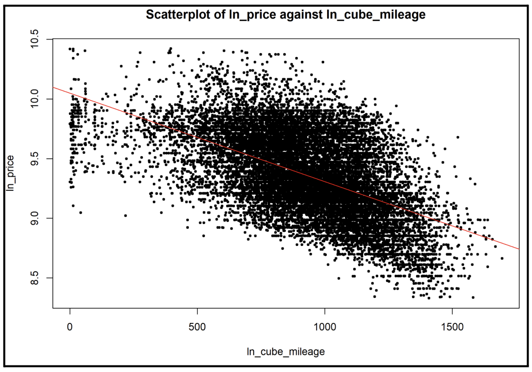
```
>t.test(car_data[car_data$fuelType=="Diesel",10],car_data[car_data$fuelType
== "Petrol",10], var.equal = F)

      Welch Two Sample t-test
```

```
data:         car_data[car_data$fuelType     ==     "Diesel",     10]     and
car_data[car_data$fuelType == "Petrol", 10]
t = 25.286, df = 9446.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1343417 0.1569207
sample estimates:
mean of x mean of y
 9.460573  9.314942
```

By the above result, the t-test returns a p-value of 2.2e^-16, which is less than 0.05 at the significance level of 0.05. Thus, H_0 is rejected and we conclude that the means are different for the two samples.

4.1.4 Relation between ln_price and mileage

For section 4.1.4, we want to determine whether the resale prices of cars depend on the existing mileage of the car itself. We performed a simple linear regression model between *ln_price* and *ln_cube_mileage*. The plot below illustrates the distribution of *ln_price* against *ln_cube_mileage*



By visualisation, it can be seen from the graph *ln_cube_mileage* linearly decreases with *ln_price* by using the red coloured linear regression line.

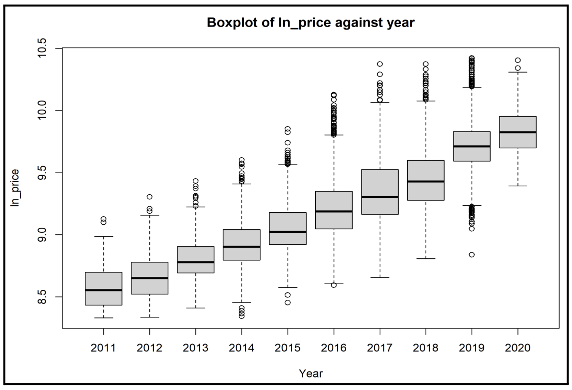
```
Call:
lm(formula = car_data$ln_price ~ car_data$ln_cube_mileage)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96978 -0.20108 -0.02122  0.18657  0.99640

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.005e+01  8.097e-03 1240.97   <2e-16 ***
car_data$ln_cube_mileage -7.411e-04  8.428e-06  -87.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2825 on 17139 degrees of freedom
Multiple R-squared:  0.3109,   Adjusted R-squared:  0.3109
F-statistic:  7733 on 1 and 17139 DF,  p-value: < 2.2e-16
```

The regression model provided a p-value of <2e-16, which is smaller than 0.05 at the significance level of 0.05. Thus, the null hypothesis (*ln_price* is related to *mileage*) is rejected.

Moreover, we identified that the R-squared for this model is 0.3109, concluding that the mileage can only explain 31.09% of the variation in *ln_price*. Therefore, the model is not a good fit to the data.

4.1.5 Relation between ln_price and year

For section 4.1.5, we want to determine whether the resale prices of cars depend on the year. We performed an analysis of variance (1-way ANOVA) test to determine whether *ln_price* has the same mean values for all the *years*. The plot below illustrates the distribution of *ln_price* from 2011 to 2020.



By visualisation, the spread of *ln_price* is not the same for all the years. It is obvious that the means of *ln_price* increases linearly with years.

By hypothesis testing (ANOVA test):

H_0: $\mu_{2011} = \mu_{2012} = \mu_{2013} = \mu_{2014} = \mu_{2015} = \mu_{2016} = \mu_{2017} = \mu_{2018} = \mu_{2019} = \mu_{2020}$

H_1: not all μ_i are equal

```
                          Df Sum Sq Mean Sq F value Pr(>F)
factor(car_data$year)      9 1016.5  112.94    1998 <2e-16 ***
Residuals              17131  968.3    0.06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA test returns a p-value of <2e-16, which is less than 0.05 at the significance level of 0.05. Thus, H_0 is rejected, and we conclude that the prices of cars are independent of the years.

Since the null hypothesis is rejected, we use pairwise.t.test() to determine which transmissions have the same means.

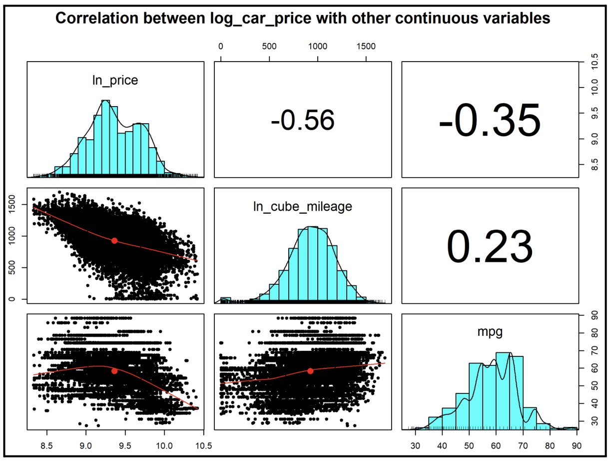
```
       Pairwise comparisons using t tests with pooled SD

data:  car_data$ln_price and car_data$year

     2011    2012    2013    2014    2015    2016    2017    2018    2019
2012 0.016   -       -       -       -       -       -       -       -
2013 5.1e-13 1.8e-06 -       -       -       -       -       -       -
2014 < 2e-16 < 2e-16 < 2e-16 -       -       -       -       -       -
2015 < 2e-16 < 2e-16 < 2e-16 < 2e-16 -       -       -       -       -
2016 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 -       -       -       -
2017 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 -       -       -
2018 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 -       -
2019 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 -
```

```
2020 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 3.9e-12
```

By the above result, the null hypothesis is rejected for all the years' pairs at significance level 0.05. Hence, we conclude that $\mu_{2011} \neq \mu_{2012} \neq \mu_{2013} \neq \mu_{2014} \neq \mu_{2015} \neq \mu_{2016} \neq \mu_{2017} \neq \mu_{2018} \neq \mu_{2019} \neq \mu_{2020}$

4.2 Linear Regression models and correlation

4.2.1 Correlations between ln_price and other continuous variables



From the matrix scatter plot,

- *mpg* is positively correlated to *ln_cube_mileage* with Pearson correlation coefficient = *0.23*. (It is a weak linear relationship)
- *ln_price* is negatively correlated to *mpg* with Pearson correlation coefficient = *-0.35*. (It is a weak linear relationship)
- *ln_price* is also negatively correlated to *ln_cube_mileage* with Pearson correlation coefficient = *-0.56*. (It is a moderate linear relationship)

Specifically, it was noteworthy to observe that the variable *ln_cube_mileage* has the highest correlation with the variable of interest (*ln_price*).

4.2.2 Single Linear Regression models

From the above section, we encountered several variables that are correlated to our variable of interest (*ln_price*), thus we wish to perform simple linear regression analysis for determining which of the three variables could model *ln_price* in a linear fashion.

Suggested model:

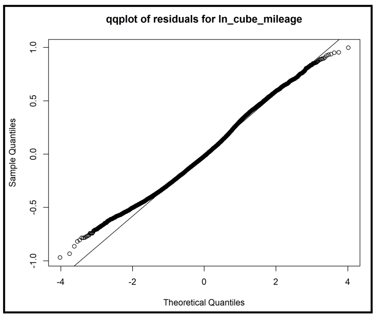
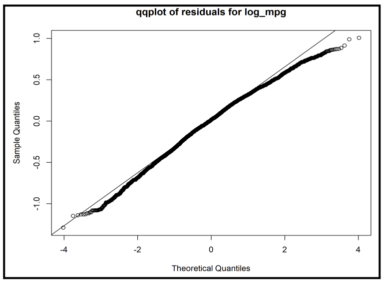
$ln(price) = \beta_0 + \beta_1 * X + \epsilon$

β_0 denotes y-intercept of linear model

β_1 denotes coefficient of predictor variable X

X denotes an one of *ln_cube_mileage*, *mpg*

ϵ denotes random error / residual term of linear model

Variable (X)	Fitted Model	p-value	R-squared	qqplot of residuals
ln_cube_mileage	$Y = 10.05 - 7.41 * 10^{-4} X$	<2.2e-16	0.3109	
mpg	$Y = 10.09 - 0.01X$	<2.2e-16	0.124	

Among the two variables, *ln_cube_mileage* is a better performance measure to model *ln_price* because it has a R-squared value of 0.3109, which means the proposed linear model can explain about 31% of variation in outcome variable *ln_price*.

4.2.3 Multiple Linear Regression model

We attempt to build a multiple linear model for *ln_price* using *mpg* and *ln_cube_mileage*. We then use backward elimination methods to select the most appropriate model. The fitted model is:

$$ln(price) = 10.47 - 8.3 * 10^{-3} mpg - 6.7 * 10^{-4} ln(mileage)^3$$

The fitted model has an R-squared value of 0.3635, which means the proposed linear model can explain about 37% of variation in outcome variable *ln_price*.

```
> summary(multipleModel)

Call:
lm(formula = car_data$ln_price ~ car_data$mpg + car_data$ln_cube_mileage)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98611 -0.17692  0.00024  0.18302  0.94252

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.047e+01  1.365e-02  766.94   <2e-16 ***
car_data$mpg             -8.396e-03  2.231e-04  -37.64   <2e-16 ***
```

```
car_data$ln_cube_mileage -6.686e-04  8.326e-06  -80.30   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2715 on 17138 degrees of freedom
Multiple R-squared:  0.3635,     Adjusted R-squared:  0.3635
F-statistic:  4894 on 2 and 17138 DF,  p-value: < 2.2e-16

> step(multipleModel, direction='backward')
Start:  AIC=-44693.96
car_data$ln_price ~ car_data$mpg + car_data$ln_cube_mileage

                           Df Sum of Sq    RSS    AIC
<none>                                  1263.3 -44694
- car_data$mpg              1    104.44 1367.7 -43334
- car_data$ln_cube_mileage  1    475.31 1738.6 -39222

Call:
lm(formula = car_data$ln_price ~ car_data$mpg + car_data$ln_cube_mileage)

Coefficients:
             (Intercept)               car_data$mpg  car_data$ln_cube_mileage
              10.4700647                 -0.0083962                -0.0006686
```
We conclude that both variables are significant measures to model *ln_price* since they are not eliminated by the Stepwise Algorithm.

5. Conclusion and Discussion

The automotive industry, Ford, is renowned for innovation and diverse revenue streams. We hope to provide some insights on their car pricing from a quantitative perspective. By analysing the Ford Car Price dataset, this report aims to understand the relationship between variables such as mileage, model, year, mpg, transmission, fuel type and price. Insights gained can inform strategic decisions, guiding Ford's market approach.

We conclude that:

- The geometric mean of a car's price depends on whether the fuel type is petrol or diesel.
- The geometric mean of a car's price depends on the transmission type of the car.
- The geometric mean of a car's price depends on the model of the car.
- The geometric mean of a car's price depends on the production year of the car.
- A car's mileage does not affect its car price.
- A car's mpg does not affect its car price.

We see that the mileage and mpg can be used to model the price via a linear model based on the simple linear regression models. Out of the two estimators, mileage is a better performance measure to model price.

While the findings of this report are intriguing, it's important to acknowledge that this analysis is derived from limited data available online for Ford cars. Additionally, with the evolution of data collection methods, Ford may have developed more intricate performance metrics than those examined here. A more comprehensive and in-depth analysis of the Ford car dataset, employing advanced analytical techniques, would be necessary to establish a more robust understanding of the relationship between various factors and the car price.

6. Appendix

```
#Code that we used
car_data = read.csv("/Users/Asus/Desktop/mh3511ford/ford.csv")
ori_copy = read.csv("/Users/Asus/Desktop/mh3511ford/ford.csv")
library(psych)
library(dplyr)

#data inspection
head(car_data)
car_price = car_data$price
summary(car_data$price)
hist(car_price , main = "Histogram of Price", xlab="price")
boxplot(car_price, main = "Boxplot of car_price")

#3.1.1 main variable of interest: price
#transforming car_data$price
log_car_price <- log(car_data$price)
hist(log_car_price, main = "Histogram of ln_price", xlab="ln_price")
boxplot(log_car_price, main = "Boxplot of ln_price", ylab="ln_price")

# Remove outliers
Q1 = quantile(log_car_price,0.25)
Q3 = quantile(log_car_price,0.75)
IQR = Q3-Q1
filtered_log_car_price = log_car_price[log_car_price >= Q1 -IQR*1.5 &  log_car_price <= Q3 + IQR*1.5]
hist(filtered_log_car_price, main = "Histogram of ln_price", xlab="ln_price")
boxplot(filtered_log_car_price, main = "Boxplot of ln_price", ylab="ln_price")
ln_price = filtered_log_car_price
summary(ln_price)
qqnorm(filtered_log_car_price)
qqline(filtered_log_car_price, col="blue")

# Add a ln_price column and eliminate outliers
filtered_car_data = car_data[log_car_price >= Q1 -IQR*1.5 &  log_car_price <= Q3 + IQR*1.5,]
filtered_car_data$ln_price = log(filtered_car_data$price)
head(filtered_car_data)
length(filtered_log_car_price)
length(filtered_car_data)
```

```
outliers_percentage = 1 - nrow(filtered_car_data)/nrow(car_data)
outliers_percentage
car_data = filtered_car_data

#3.2.1 model
str(car_data$model)
model_table <- aggregate(car_data$price,list(car_data$model),FUN=length)
colnames(model_table) <- c("Model","Count")
mode = model_table[model_table$Count == max(model_table$Count),]

par(mar = c(5, 9, 4, 2))  # Set margins: bottom, left, top, right
barplot(model_table$Count,
        names=model_table$Model,
        horiz = T,
        cex.names = 0.75,
        main="Number of Cars by Model (Before)",
        xlab="Count",
        las=2)

#after
models = model_table$Model[model_table$Count >= 200]
filtered = car_data[car_data$model %in% models ,]
model_table = aggregate(filtered$price, list(filtered$model), FUN=length)
colnames(model_table) = c("Model", "Count")
barplot(model_table$Count,
        names=model_table$Model,
        horiz = T,
        cex.names = 0.75,
        main="Number of Cars by Model (After)",
        xlab="Count",
        las=2)
nrow(car_data)-nrow(filtered)
1-nrow(filtered)/nrow(car_data)
car_data = filtered

#3.2.2 transmission
str(car_data)
transmission_table = aggregate(car_data$price, list(car_data$transmission), FUN=length)
colnames(transmission_table) = c("Transmission", "Count")
barplot(transmission_table$Count,
```

```
        names=transmission_table$Transmission,
        horiz = T,
        cex.names = 0.75,
        main="Cars by Transmission",
        xlab="Count",
        las=2)

#3.2.3 fuelType
str(car_data$fuelType)
fuel_type_table = aggregate(car_data$price, list(car_data$fuelType), FUN=length)
colnames(fuel_type_table) = c("Fuel_type", "Count")
fuel_type_table
barplot(fuel_type_table$Count,
        names=fuel_type_table$Fuel_type,
        horiz = T,
        cex.names = 0.75,
        main="Cars by fuelType (Before)",
        xlab="Count",
        las=2)

#after
filtered = car_data[car_data$fuelType == "Petrol" | car_data$fuelType == "Diesel",]
fuel_type_table = aggregate(filtered$price, list(filtered$fuelType), FUN=length)
colnames(fuel_type_table) = c("Fuel_type", "Count")
fuel_type_table
barplot(fuel_type_table$Count,
        names=fuel_type_table$Fuel_type,
        horiz = T,
        cex.names = 0.75,
        main="Cars by fuelType (After)",
        xlab="Count",
        las=2)
1-nrow(filtered)/nrow(car_data)
nrow(car_data)-nrow(filtered)
car_data=filtered
head(car_data)

#3.2.4 mileage
str(car_data$mileage)
hist(car_data$mileage, main="Histogram of mileage", xlab="mileage")
```

```
boxplot(car_data$mileage, main="Boxplot of mileage", ylab="mileage")
hist(log(car_data$mileage)^3, main="Histogram of (ln_mileage)^3", xlab="(ln_mileage)^3")
boxplot(log(car_data$mileage)^3, main="Boxplot of (ln_mileage)^3", ylab="(ln_mileage)^3")
sum(car_data$mileage==0)
sum(car_data$mileage<=1000)
car_data$ln_cube_mileage = log(car_data$mileage)^3
head(car_data)

#3.2.5 mpg
head(car_data)
str(car_data$mpg)
hist(car_data$mpg, main="Histogram of mpg", xlab="mpg")
boxplot(car_data$mpg, main="Boxplot of mpg", ylab="mpg")

# 3.2.6 year
str(car_data$year)
car_data = car_data[car_data$year != 2060,]
year_table = aggregate(car_data$year, list(car_data$year), FUN=length)
colnames(year_table) = c("Year", "Count")
year_table
barplot(year_table$Count,
      names=year_table$Year,
      horiz = T,
      cex.names = 0.75,
      main="Cars by Year (Before)",
      xlab="Count",
      las=2)
years = year_table$Year[year_table$Count >= 30]
filtered = car_data[car_data$year %in% years,]
nrow(car_data) - nrow(filtered)
1-nrow(filtered)/nrow(car_data)

year_table = aggregate(filtered$year, list(filtered$year), FUN=length)
colnames(year_table) = c("Year", "Count")
year_table
barplot(year_table$Count,
      names=year_table$Year,
      horiz = T,
      cex.names = 0.75,
      main="Cars by Year (After)",
```

```
        xlab="Count",
        las=2)
car_data = filtered

#4.1.1 ln_price vs model
boxplot(car_data$ln_price~car_data$model, car_data, xlab="model", ylab="ln_price", main='Boxplot of ln_price against car models')
summary(aov(car_data$ln_price~factor(car_data$model)))
pairwise.t.test(car_data$ln_price, car_data$model, p.adjust.method="none")

#4.1.2 ln_price vs transmission
boxplot(car_data$ln_price~car_data$transmission, car_data, xlab="Transmission", ylab="ln_price", main="Boxplot of ln_price against transmission")
aov(car_data$ln_price~factor(car_data$transmission))
summary(aov(car_data$ln_price~factor(car_data$transmission)))
pairwise.t.test(car_data$ln_price, car_data$transmission, p.adjust.method="none")

#4.1.3 ln_price vs fuelType
boxplot(car_data$ln_price~car_data$fuelType, car_data, xlab="fuelType", ylab="ln_price", main="Boxplot of ln_price against fuelType")
var.test(car_data[car_data$fuelType == "Diesel",10],car_data[car_data$fuelType == "Petrol",10])
t.test(car_data[car_data$fuelType == "Diesel",10],car_data[car_data$fuelType == "Petrol",10], var.equal = F)

#4.1.4 ln_price vs ln_cube_mileage
plot(car_data$ln_price~car_data$ln_cube_mileage, xlab="ln_cube_mileage", ylab="ln_price", main="Scatterplot of ln_price against ln_cube_mileage", pch=19, cex=0.5)
abline(lm(car_data$ln_price~car_data$ln_cube_mileage), col='red')
lmodel = lm(car_data$ln_price~car_data$ln_cube_mileage)
summary(lmodel)

#4.1.5 ln_price vs year
boxplot(car_data$ln_price~car_data$year, car_data, xlab="Year", ylab="ln_price", main='Boxplot of ln_price against year')
aov(car_data$ln_price~factor(car_data$year))
summary(aov(car_data$ln_price~factor(car_data$year)))
pairwise.t.test(car_data$ln_price, car_data$year, p.adjust.method="none")

#4.2.1 correlation matrix with scatterplots
```

```
pairs.panels(car_data[,c(10,11,8)],main='Correlation between ln_price with other continuous variables')

#4.2.2 single linear regression
model1 = lm(car_data$ln_price ~ car_data$ln_cube_mileage)
summary(model1)
qqnorm(model1$residuals, main='qqplot of residuals for ln_cube_mileage')
qqline(model1$residuals)
model2 = lm(car_data$ln_price ~ car_data$mpg)
summary(model2)
qqnorm(model2$residuals, main='qqplot of residuals for log_mpg')
qqline(model2$residuals)

#4.2.3 multiple linear regression
multipleModel = lm(car_data$ln_price~ car_data$mpg + car_data$ln_cube_mileage)
summary(multipleModel)
step(multipleModel, direction='backward')
```

7. References

Wayland M. 2024. Ford reports 7.1% increase in U.S. new vehicle sales as industry marks best year since 2019. Consumer News and Business Channel. [accessed 2024 Apr 4]. https://www.cnbc.com/2024/01/04/ford-reports-7point1percent-increase-in-us-vehicle-sales.html.

Statista. 2024. Ford Motor Company's revenue from FY 2008 to FY 2023 [accessed 2024 Apr 4] . https://www.statista.com/statistics/267305/total-revenue-of-ford/.

Tech Xplore. 2023. Ford reports higher US auto sales amid strong demand. [accessed 2024 Apr 4]. https://techxplore.com/news/2023-10-ford-higher-auto-sales-strong.html.

The Business Times. 2024. Ford's 2023 US sales rise 7.1% on pickup truck, SUV demand. [accessed 2024 Apr 4]. https://www.businesstimes.com.sg/companies-markets/fords-2023-us-sales-rise-71-pickup-truck-suv-demand.