ASSIGNMENT COVER SHEET

Research School of Finance, Actuarial
Studies and Statistics
ANU College of Business and
Economics
Australian National University
Canberra ACT 0200 Australia
www.anu.edu.au

This cover sheet should be attached to the front of your assignment, which must be submitted online using Turnitin.

Student ID For group assignments, list each student's ID	U7446204		
Course Code	STAT6038		
Course Name	Regression Modelling		
Assignment number	Assignment2		
Assignment Topic	Analysing Used Car Prices with Simple Linear Regression		
Lecturer	Insha Ullah		
Tutor			
Tutorial (day and time)			
Word count	2275	Due Date	18 May 2023 - 15:00
Date Submitted	17 May 2023	Extension Granted	

I declare that this work:

☑upholds the principles of academic integrity, as defined in the ANU Policy*: Code of Practice for Student Academic Integrity*;

☑is original, except where collaboration (for example, group work) has been authorised in writing by the course convener in the course outline and/or Wattle site;

☑is produced for the purposes of this assessment task and has not been submitted for assessment in any other context, except where authorised in writing by the course convener;

☑gives appropriate acknowledgement of the ideas, scholarship and intellectual property of others insofar as these have been used;

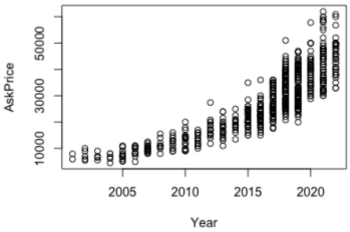
☑in no part involves copying, cheating, collusion, fabrication, plagiarism or recycling.

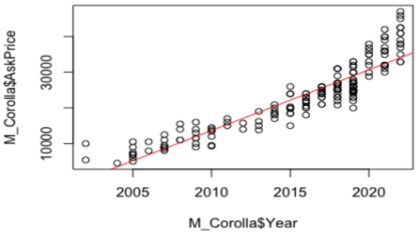
Initials For group assignments, each student must initial (digital is acceptable provided all participants have confirmed)	Xiaoxu Kuang U7446204

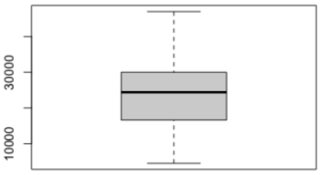
a)

Based on the scatter plot of Year versus AskPrice, it is a short guess that Year shows a positive correlation with Price, which means that as the year of manufacture increases, so does AskPrice.

```
## Model
##   Camry Corolla
##     638     180

length(unique(Model))

## [1] 2
```

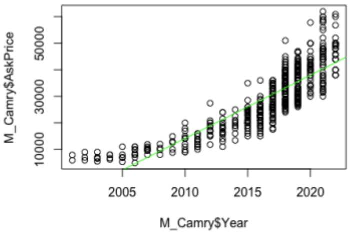


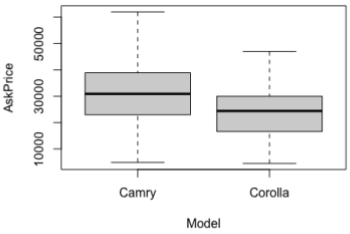




Based on a preliminary analysis of the model data, the model data is of the category data type and is only available in Corolla and Camry.

The data for Corolla's year of manufacture and AskPrice were extracted and combined, and the data under this category were sorted and plotted in a plot and box plot. year and AskPrice were found to be linearly related in the long term, and Corolla's AskPrice was found to be positively correlated. The mean is close to 23711.



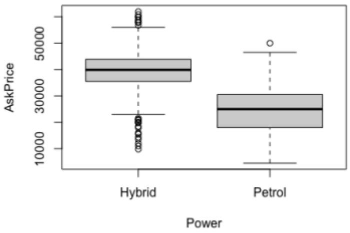


The data for "year of manufacture" and AskPrice were extracted from Model Camry and merged, and the data under this category were collated and plotted and box plotted. year and AskPrice were found to be positively correlated in the long run but not simply linear, and further analysis of the data was required, initially The initial analysis seems to have a quadratic-like correlation. The mean of Camry's Askprice is close to 30425.

The average AskPrice for Camry is higher than for cars with the Corolla model.

b)

Based on the initial processing of the power data and the analysis, it was found that the variable power type was divided into 2 categories, Hybrid and Petrol.



As AskPrice is a discrete data type and Power data is a category data type. It was therefore decided to extract AskPrice for both power type cars for analysis and to boxplot AskPrice against Power at the same time.

According to the summary() data, the average AskPrice for a car with power type Hybrid is 39126, while the average AskPrice for a car with power type Petrol is 24152. Hybrid type cars seem to be more popular in the market.

```
summary(P_Hybrid_P)

##     AskPrice
##  Min.   : 9800
##  1st Qu.:35597
##  Median :39884
##  Mean   :39126
##  3rd Qu.:43904
##  Max.   :61970

summary(P_Petrol_P)

##     AskPrice
##  Min.   : 4500
##  1st Qu.:17990
##  Median :24990
##  Mean   :24152
##  3rd Qu.:30500
##  Max.   :49990
```

c)

```
table (Transmission)

## Transmission
## Automatic
##       818

length(unique(Transmission))

## [1] 1
```

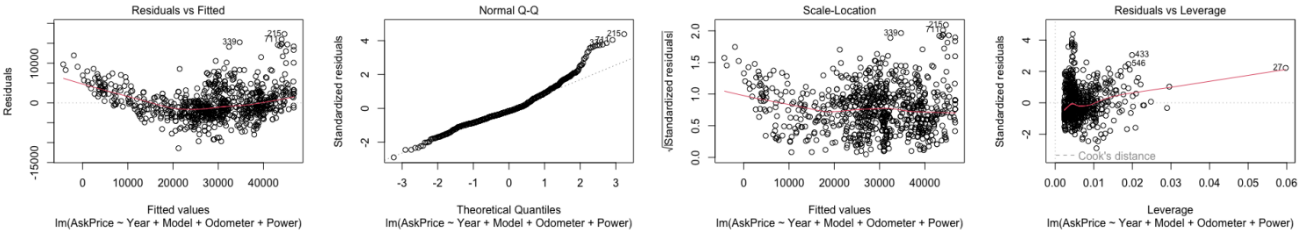
"transmission" is NOT useful as a predictor in a multiple regression model.

Due to the category of Transmission is only one type- Automatic, which means this variable does not show its effect when analyzing the AskPrice because all AskPrices are in the same category- Automatic.

d)

(i)

According to the result of part(c), we should eliminate the "transmission" from the MLR model if we analyze the AskPrice.



Analyzing the residuals versus the fitted value plot, if the assumption of linearity is satisfied, then the residuals should be randomly distributed around the horizontal line without any systematic trend or pattern. In the graph, it is clear that there is a near fan-shaped distribution here and therefore the assumption of linearity is not satisfied.

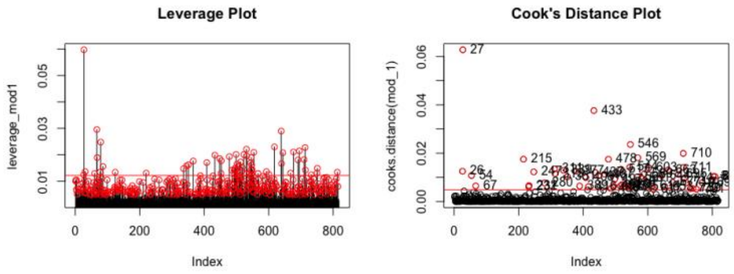
There may be a non-linear relationship because the curve grows upwards and has a concave trend.

Furthermore, the graph is used to test the 'Constance variance' assumption, which is clearly violated by this trending and concentrated distribution of data.

Besides, outliers are 215, 339, 711 in the observed data.

With regard to normal Q-Q plots, if the assumption of normality of the residuals is met, then the residuals should fall roughly on a straight line. In this plot, significant bends or deviations are seen, with outliers of 215, 339 and 711 respectively.

```
# Create a bar plot of the leverages for each observation
leverage_mod1 <- hatvalues (mod_1)
n <- nrow (subset_data1)
p <- length (coef (mod_1))
highlev <- which (leverage_mod1 >= 4/(n-p))
plot(leverage_mod1, type = "h", main = "Leverage Plot")
abline( h = 2*mean(leverage_mod1) , col ="red")
points(highlev, leverage_mod1[highlev], col = "red", pch = 1)
```



If the leverage value of an observation is set close to or above 4/(n - p), where n is the number of samples and p is the number of independent variables, then it may be a high leverage point. Finding too many high leverage values according to this criterion to the extent that it is difficult to distinguish them indicates that the model is not appropriate. It is therefore necessary to look at the outliers with the help of the leverage plot that comes with the plot() function.

These strong impact points are observations that have a strong influence on the estimation of the model parameters. Based on the two plots, it can be seen that the outliers are 27, 433, 546.

```
# Create a bar plot of Cook's distances for each observation
plot(cooks.distance(mod_1), main = "Cook's Distance Plot")
abline(h = 4/length(cooks.distance(mod_1)), col= "red")
cooksd_mod1<-cooks.distance(mod_1)
cooks_abline_mod1 <- 4/length(cooksd_mod1)
cooks_high_mod1 <- which(cooksd_mod1 > cooks_abline_mod1)
points(cooks_high_mod1, cooksd_mod1[cooks_high_mod1], pch = 1, col = "red")
text(cooks_high_mod1, cooksd_mod1[cooks_high_mod1], cooks_high_mod1, pos = 4)
```

In the figure, if the Cook distance of an observation is greater than 1 or 4/(n-p), then it may be a strong influence point. If the image shows that too many observations are strong influence points according to this criterion, perhaps the model is not suitable or perfect. Particularly obvious outliers are 27, 433, 546.

(ii)

All outliers that may affect the multiple regression analysis have been named in the above (like: point 27, 215, 339,433, 546, 711) and should be removed before the remaining data is analysed.

e)

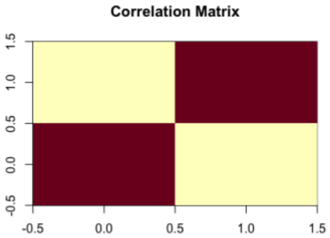
As both Model and Power are category data types, there is a multicollinearity relationship between the two types before them. This should probably be eliminated when analyzing the multicollinearity relationship.

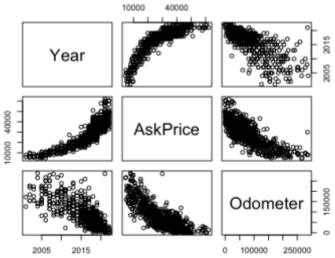
```
subset_data2<- carsales_2[,-c(2,5,6)]
pairs(subset_data2)
```

Based on the plot drawn by pairs(), it is guessed that AskPrice has a more obvious relationship with the Year data after square ("polynomial terms"), while the relationship between the Odometer data and the AskPrice data is unclear and therefore needs to be "transformation". It is also necessary to test for a multicollinearity relationship between Year and Odometer.

```
cor_matrix <- cor(subset_data1[, c("Year","Odometer")])
inv_cor_matrix <- solve(cor_matrix)
VIF_values <- diag(inv_cor_matrix)
VIF_values

##     Year Odometer
## 2.420274 2.420274
```

According to the "rule of dumb" of VIF, VIF> 10 should be rejected, the result here is 2.420274 there seems to be no multicollinearity relationship in the model, so we can "transformation" the data of Year and Odometer at the same time, and finally find the right model.





f)

Based on the information of lecture, if a polynomial term of a given order is retained, then all related terms of lower order are also retained in the model, and we don't really consider models with order higher than 3.

```
model1<-lm(AskPrice~ Year+ I(log(Year)) + Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model1,main = "1")

model1_1<-lm(AskPrice~ I(log(Year)) + Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model1_1,main = "1_1")

model2<-lm(AskPrice~ Year + Model + Odometer+ I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model2, main = "2")

model2_2<-lm(AskPrice~ Year + Model +  I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model2_2,main="2_2")

model3<-lm((log(AskPrice))~Year + Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model3, main="3")

model4<-lm(AskPrice~  I(log(Year)) + Model +  I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model4, main="4")

model4_2<-lm(AskPrice~ Year+ I(log(Year)) + Model + Odometer+ I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model4_2, main = "4_2")
```

```
model5<-lm(AskPrice ~Year + I(Year^2) +Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model5, main= "5")

model6<-lm(AskPrice ~Year + I(Year^2)+ Model + I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model6 , main= "6")

model7<-lm(AskPrice ~Year + I(Year^2) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model7, main = "7")

model8<-lm(AskPrice ~I(log(Year)) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model8, main = "8")

model9<-lm((AskPrice^2)~Year + Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model9, main= "9")

model10<-lm(sqrt(AskPrice)~ Year + Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model10, main="10")

model11<-lm(sqrt(AskPrice)~ Year + I(Year^2) +Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model11, main="11")

model12<-lm(sqrt(AskPrice)~ Year + I(Year^2)+ Model + I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model12, main="12")

model13<-lm(sqrt(AskPrice)~ Year + I(Year^2) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model13, main="13")

model14<-lm(sqrt(AskPrice)~ (log(Year)) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model14, main="14")

model15<-lm((AskPrice^2)~ Year + I(Year^2) +Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model15, main="15")

model16<-lm((AskPrice^2)~ Year + I(Year^2)+ Model + I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model16, main="16")

model17<-lm((AskPrice^2)~ Year + I(Year^2) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model17, main="17")
```

```
model18<-lm((AskPrice^2)~ (log(Year)) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model18, main="18")

model19<-lm((log(AskPrice))~Year + I(Year^2) +Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model19, main="19")

model20<-lm((log(AskPrice))~Year + I(Year^2)+ Model + I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model20, main="20")

model21<-lm((log(AskPrice))~Year + I(Year^2) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model21, main="21")

model22<-lm((log(AskPrice))~(log(Year)) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model22, main="22")

summary(model1)

summary(model3)

summary(model4_2)

summary(model5)

summary(model7)

summary(model11)

summary(model12)

summary(model13)

summary(model14)

summary(model19)

summary(model20)

summary(model21)

summary(model22)
```

We used the "Residuals VS Fitted" image to select the more suitable candidates. Candidates selected according to the chart, respectively mod: 1,3,4_2,5,7,11,12,13,14,19,20,21,22

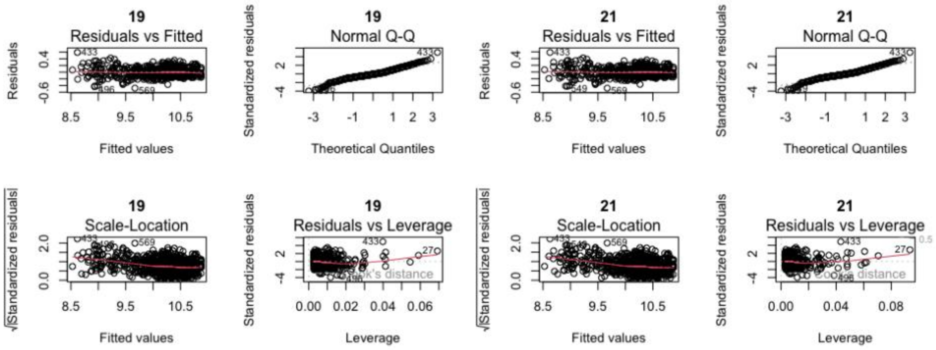
(i)

Based on the value of "Adjusted R-squared" calculated by the summary() function, we pick the largest models, which should be model19(Adjusted R-squared=0.9384) and model21(Adjusted R-squared=0.9384).

model19<-lm((log(AskPrice))~Year + I(Year^2) +Model + Odometer + Power)

model21<-lm((log(AskPrice))~Year + I(Year^2) +Model + Odometer+I(Odometer^2) + Power)

The MLR reflected by these two mods is shown above.



```
summary(model19)
## Call:
## lm(formula = (log(AskPrice)) ~ Year + I(Year^2) + Model + Odometer +
##     Power)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47831 -0.07231 -0.01400  0.06262  0.59127
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.976e+03  7.516e+02  -3.960 8.16e-05 ***
## Year          2.893e+00  7.469e-01   3.873 0.000116 ***
## I(Year^2)    -7.001e-04  1.855e-04  -3.773 0.000173 ***
## ModelCorolla -1.426e-01  1.048e-02 -13.610  < 2e-16 ***
## Odometer     -2.123e-06  1.178e-07 -18.030  < 2e-16 ***
## PowerPetrol  -2.107e-01  9.988e-03 -21.099  < 2e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.1212 on 812 degrees of freedom
## Multiple R-squared:  0.9387, Adjusted R-squared:  0.9384
## F-statistic:  2488 on 5 and 812 DF,  p-value: < 2.2e-16

summary(model21)
## Call:
## lm(formula = (log(AskPrice)) ~ Year + I(Year^2) + Model + Odometer +
##     I(Odometer^2) + Power)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47820 -0.07101 -0.01267  0.06159  0.58242
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.506e+03  8.417e+02  -2.978  0.00299 **
## Year           2.426e+00  8.365e-01   2.900  0.00383 **
## I(Year^2)     -5.840e-04  2.078e-04  -2.810  0.00507 **
## ModelCorolla  -1.423e-01  1.048e-02 -13.588   < 2e-16 ***
## Odometer      -1.800e-06  2.866e-07  -6.279 5.54e-10 ***
## I(Odometer^2) -1.319e-12  1.066e-12  -1.238  0.21606
## PowerPetrol   -2.115e-01  1.000e-02 -21.142  < 2e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.1212 on 811 degrees of freedom
## Multiple R-squared:  0.9388, Adjusted R-squared:  0.9384
## F-statistic:  2075 on 6 and 811 DF,  p-value: < 2.2e-16
```

(ii)

The F-statistic is a statistic used to test whether the linear regression model is set appropriately. A larger F-statistic indicates a better overall fit of the regression model.

The p-value is the probability value associated with the F-statistic and is used to assess the significance of the F-statistic. Typically, we use the p-value for hypothesis testing. In the case of MLR, the summary() function helps to obtain the results of the significance test for each independent variable, if the p-value is <0.05, there is reason to believe that the variables are not statistically significant, which can help to choose whether the model is reasonable or not.

In summary, the F-statistic and the p-value together provide information on whether the overall goodness of fit of the linear regression model is significant. Both the larger F-statistic and the smaller p-value imply that the model as a whole is significant, i.e. that the combination of independent variables explains the dependent variable in a statistically significant way.

(iii)

Based on the data from the summary() function, it is known that:

In model19, which measures the significance of Year and Year^2, their p-value is 0.000116, 0.000173 respectively, both less than 0.05 (the default significance level), and for Odometer the significance of p-value is < 2e-16, which is also insignificant.

In model21, which measures the significance of Year and Year^2, their p-values are 0.003830.00507, respectively, with p-value > 0.05 (default significant level) for Year^2, which is considered significant as a variable in the MLR case, for significance of Odometer and Odometer^2, the p-values are 5.54e-10, 0.21606, respectively, with p-value>0.05 (default significant level) for Odometer^2, which is considered significant as a variable in the MLR case.

# Therefore, in the case where the adjusted R- squared of both models is a maximum of 0.9384, the model21 with the more significant variable should be chosen

g)

According to part f) we have chosen "model21<-lm((log(AskPrice))~Year + I(Year^2) + Model + Odometer + I(Odometer^2) + Power)". The data first needs to be acquired and analyzed using the predicted data, with the significant level using the default of 5%. In addition, since the dependent variable "AskPrice" is transformed by log() in the model, the prediction value should also be manipulated using the exp() function.

```
test_data<-read.csv("test_data.csv")
Prediction_AskPirce <- predict(model21, newdata = data.frame(test_data), interval = "prediction", level =
0.95)
exp(Prediction_AskPirce)

##        fit      lwr      upr
## 1 16237.07 12785.70 20620.11
## 2 26714.30 21039.07 33920.40
## 3 31909.64 25144.09 40495.60
## 4 39430.24 31038.76 50090.40
## 5 26830.34 21142.60 34048.19
## 6 29291.05 23081.92 37170.45
## 7 44896.87 35359.71 57006.39
## 8 47345.11 37259.90 60160.10
```

The result generates a price range for AskPrice, given the mod21 we set, and the price range is the predicted value obtained based on the new data.

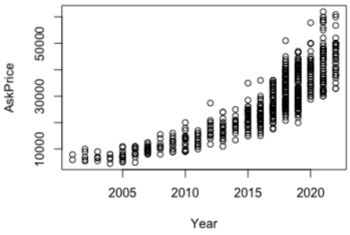
Appendix

```
carsales_2 <- read.csv("carsales.csv", header = T)
attach(carsales_2)
head(carsales_2)

##   Year   Model AskPrice Odometer Transmission  Power
## 1 2021 Corolla    35990    28929    Automatic Hybrid
## 2 2022 Corolla    37888    14020    Automatic Hybrid
## 3 2013   Camry    23950   128870    Automatic Hybrid
## 4 2022 Corolla    42499     2400    Automatic Hybrid
## 5 2015   Camry    22975    99143    Automatic Petrol
## 6 2022   Camry    49950        9    Automatic Hybrid
```

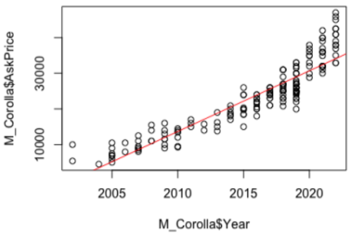
##Analysing Used Car Prices with Simple Linear Regression

#a)

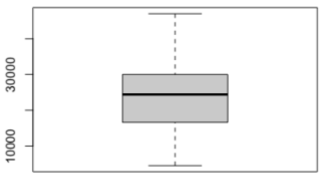
```
plot(Year, AskPrice)
```



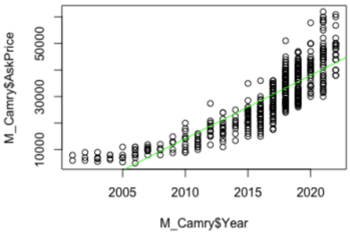
```
# Based on the scatter plot of Year versus AskPrice, it is a short guess that Year shows a positive correlation with Price, which means that as the year of manufacure increases, so does AskPrice.
table (Model)

## Model
##   Camry Corolla
##     638     180

length(unique(Model))

## [1] 2

# Based on a preliminary analysis of the model data, the model data is of the category data type and is only available in Corolla and Camry.
M_Corolla_Y<- subset(carsales_2, Model=="Corolla", select = Year)
M_Corolla_P<- subset(carsales_2, Model=="Corolla", select = AskPrice)
M_Corolla<-cbind(M_Corolla_Y, M_Corolla_P)
plot(M_Corolla$AskPrice~ M_Corolla$Year)
abline(lm(M_Corolla$AskPrice~M_Corolla$Year), col="red")
```



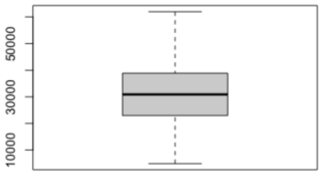
```
boxplot(M_Corolla_P, M_Corolla_Y)
```



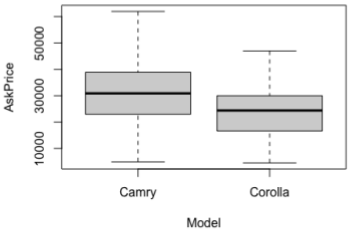
```
summary(M_Corolla)

##       Year          AskPrice
##  Min.   :2002   Min.   : 4500
##  1st Qu.:2013   1st Qu.:16795
##  Median :2018   Median :24395
##  Mean   :2016   Mean   :23711
##  3rd Qu.:2019   3rd Qu.:29996
##  Max.   :2022   Max.   :46990

# The data for Corolla's year of manufacture and askPrice were extracted and combined, and the data under th
is category were sorted and plotted in a plot and box plot. year and askPrice were found to be linearly related
 in the long term, and Corolla's askprice was found to be positively correlated. The mean is close to 23711.
M_Camry_Y<- subset(carsales_2, Model=="Camry", select = Year)
M_Camry_P<- subset(carsales_2, Model=="Camry", select = AskPrice)
M_Camry<-cbind(M_Camry_Y, M_Camry_P)
plot(M_Camry$AskPrice~ M_Camry$Year)
abline(lm(M_Camry$AskPrice~M_Camry$Year), col="green")
```



```
boxplot(M_Camry_P, M_Camry_Y)
```



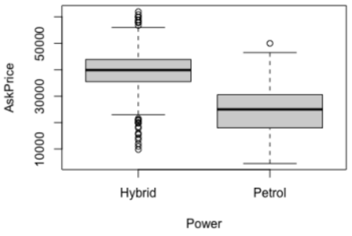
```
summary(M_Camry)

##       Year         AskPrice    
##  Min.   :2001   Min.   : 4900  
##  1st Qu.:2016   1st Qu.:22976  
##  Median :2018   Median :30888  
##  Mean   :2017   Mean   :30425  
##  3rd Qu.:2019   3rd Qu.:38886  
##  Max.   :2022   Max.   :61970  

# The data for year of manufacture and AskPrice were extracted from Model Camry and merged, and the data under this category were collated and plotted and box plotted. year and AskPrice were found to be positively correlated in the long run but not simply linear, and further analysis of the data was required, initially The initial analysis seems to have a quadratic-like correlation. The mean of Camry's Askprice is close to 30425.
boxplot(AskPrice~Model)
```



```
# The average AskPrice for Camry is higher than for cars with the Corolla model.
```

#b)

```
table (Power)

## Power
## Hybrid Petrol 
##    262    556 

length(unique(Power))

## [1] 2

# Based on the initial processing of the power data and the analysis, it was found that the variable power type was divided into 2 categories, Hybrid and Petrol.
boxplot(AskPrice~Power)
```



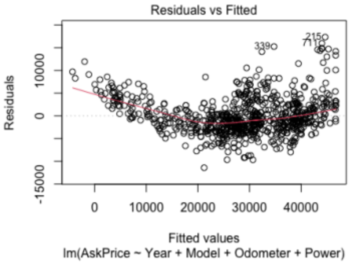
```
P_Hybrid_P<-subset(carsales_2, Power=="Hybrid", select = AskPrice)
summary(P_Hybrid_P)

##        AskPrice
##   Min.     : 9800
##   1st Qu.:35597
##   Median  :39884
##   Mean     :39126
##   3rd Qu.:43904
##   Max.     :61970

P_Petrol_P<-subset(carsales_2, Power=="Petrol", select = AskPrice)
summary(P_Petrol_P)

##        AskPrice
##   Min.     : 4500
##   1st Qu.:17990
##   Median  :24990
##   Mean     :24152
##   3rd Qu.:30500
##   Max.     :49990

# As AskPrice is a discrete data type and Power data is a category data type. It was therefore decided to extr
act AskPrice for both power type cars for analysis and to boxplot AskPrice against Power at the same time.
# According to the summary() data, the average AskPrice for a car with power type Hybrid is 39126, while the
average AskPrice for a car with power type Petrol is 24152. Hybrid type cars seem to be more popular in the
market.
```

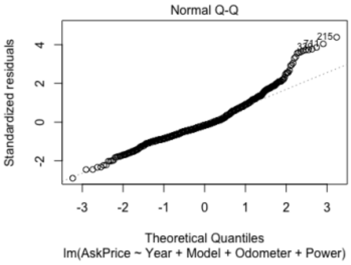
#c)

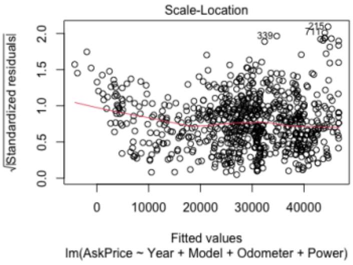
```
table (Transmission)

## Transmission
## Automatic
##          818

length(unique(Transmission))

## [1] 1

# "transmission" is NOT useful as a predictor in a multiple regression model.
# Due to the category of Transmission is only one type- Automatic, which means this variable does not show it
s effect when analyzing the AskPrice because all AskPrices are in the same category- Automatic.
```

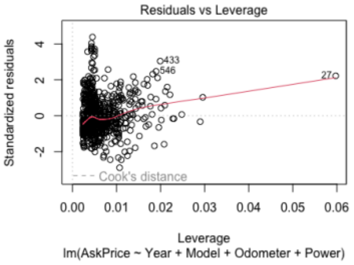
#d)

```
# i)
# According to the result of part(c), we should eliminate the "transmission" from the MLR model if we analyze the AskPrice.
names(carsales_2)

## [1] "Year"         "Model"        "AskPrice"     "Odometer"     "Transmission"
## [6] "Power"

subset_data1<- carsales_2[,-5]
mod_1<-lm(AskPrice~ Year+ Model+ Odometer+ Power)
plot(mod_1)
```



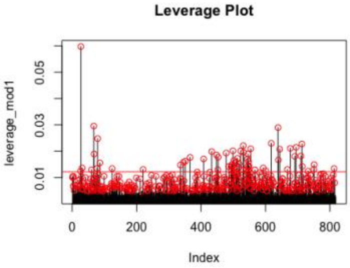




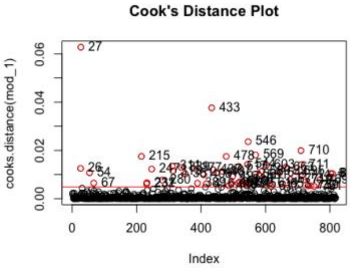


```
# Analyse the residuals versus the fitted value plot, if the assumption of linearity is satisfied, then the residuals should be randomly distributed around the horizontal line without any systematic trend or pattern. In the graph, it is clear that there is a near fan-shaped distribution here and therefore the assumption of linearity is not satisfied;
# Seeing that the curve grows upwards and has a concave trend, there may be a non-linear relationship.
# Furthermore, the graph is used to test the 'Constance variance' assumption, which is clearly violated by this trending and concentrated distribution of data.
#Besides, outliers are 215, 339, 711 in the observed data.

# With regard to normal Q-Q plots, if the assumption of normality of the residuals is met, then the residuals should fall roughly on a straight line. In this plot, significant bends or deviations are seen, with outliers of 215, 339 and 711 respectively.

# Create a bar plot of the leverages for each observation
leverage_mod1 <- hatvalues (mod_1)
n <- nrow (subset_data1)
p <- length (coef (mod_1))
highlev <- which (leverage_mod1 >= 4/(n-p))
plot(leverage_mod1, type = "h", main = "Leverage Plot")
```

```
abline( h = 2*mean(leverage_mod1) , col ="red")
points(highlev, leverage_mod1[highlev], col = "red", pch = 1)
```



```
# If the leverage value of an observation is set close to or above 4/(n - p), where n is the number of samples and p is the number of independent variables, then it may be a high leverage point. Finding too many high leverage values according to this criterion to the extent that it is difficult to distinguish them indicates that the model is not appropriate. It is therefore necessary to look at the outliers with the help of the leverage plot that comes with the plot() function.
# These strong impact points are observations that have a strong influence on the estimation of the model parameters. Based on the two plots, it can be seen that the outliers are 27, 433, 546.

# Create a bar plot of Cook's distances for each observation
plot(cooks.distance(mod_1), main = "Cook's Distance Plot")
abline(h = 4/length(cooks.distance(mod_1)), col= "red")
cooksd_mod1<-cooks.distance(mod_1)
cooks_abline_mod1 <- 4/length(cooksd_mod1)
cooks_high_mod1 <- which(cooksd_mod1 > cooks_abline_mod1)
points(cooks_high_mod1, cooksd_mod1[cooks_high_mod1], pch = 1, col = "red")
text(cooks_high_mod1, cooksd_mod1[cooks_high_mod1], cooks_high_mod1, pos = 4)
```

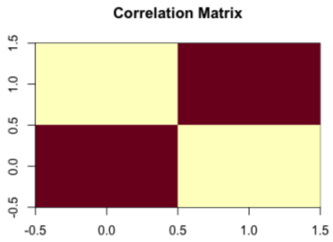


```
# The graph is used to detect the impact of each data point on the model. If the data points have a small impact on the model, their Cook's distance is small. If some data points have a large Cook's distance, they have a large influence on the model and may be outliers or outliers. Therefore, these data points can be considered for exclusion when fitting the model.
# In the figure, if the Cook distance of an observation is greater than 1 or 4/(n-p), then it may be a strong influence point. If the image shows that too many observations are strong influence points according to this criterion, perhaps the model is not suitable or perfect.Particularly obvious outliers are 27, 433, 546.
#ii)
# All outliers that may affect the multiple regression analysis have been named in the above(like: point 27, 215, 339,433, 546, 711 ) and should be removed before the remaining data is analysed.
summary(mod_1)
```

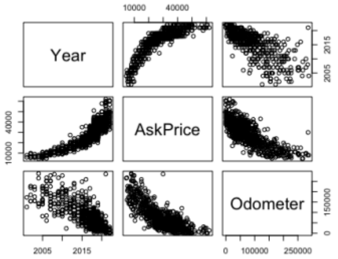
```
##
## Call:
## lm(formula = AskPrice ~ Year + Model + Odometer + Power)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11408.4  -2462.8   -643.4   2151.0  17322.3
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.625e+06  1.024e+05  -25.62   <2e-16 ***
## Year           1.321e+03  5.067e+01   26.07   <2e-16 ***
## ModelCorolla  -3.895e+03  3.413e+02  -11.41   <2e-16 ***
## Odometer      -5.558e-02  3.780e-03  -14.70   <2e-16 ***
## PowerPetrol   -8.062e+03  3.209e+02  -25.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3961 on 813 degrees of freedom
## Multiple R-squared:  0.8795, Adjusted R-squared:  0.8789
## F-statistic:  1483 on 4 and 813 DF,  p-value: < 2.2e-16
```

#e)

```
# As both Model and Power are category data types, there is a multicollinearity relationship between the two types before them. This should probably be eliminated when analysing the multicollinearity relationship.
mod_2<- lm(AskPrice~ Year+ Odometer)
mod_mat <- model.matrix(mod_2)
mod_cor <- cor(mod_mat[, -1])
image(mod_cor, main = "Correlation Matrix")
```



```
subset_data2<- carsales_2[,-c(2,5,6)]
pairs(subset_data2)
```



# Based on the plot drawn by pairs(), it is guessed that AskPrice has a more obvious relationship with the Year data after square ("polynomial terms"), while the relationship between the Odometer data and the AskPrice data is unclear and therefore needs to be "transformation". It is also necessary to test for a multicollinearity relationship between Year and Odometer.

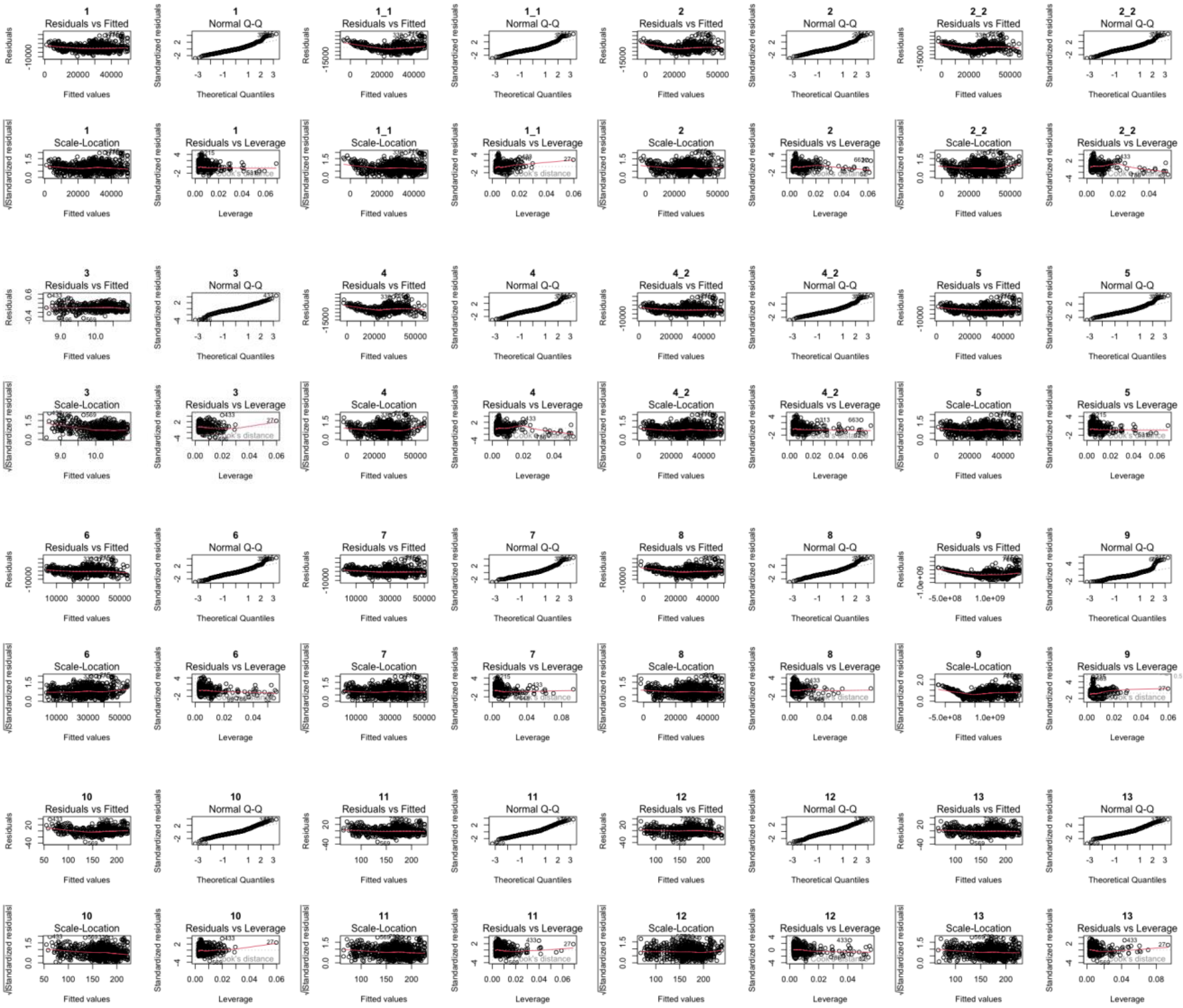
```
cor_matrix <- cor(subset_data1[, c("Year","Odometer")])
inv_cor_matrix <- solve(cor_matrix)
VIF_values <- diag(inv_cor_matrix)
VIF_values
```

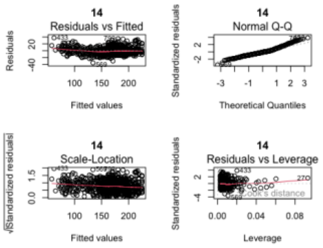
```
##        Year  Odometer
## 2.420274  2.420274
```

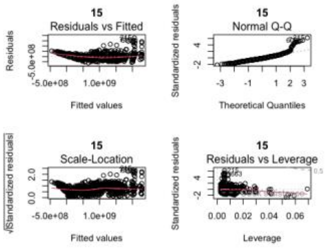
# According to the "rule of dumb" of VIF, VIF> 10 should be rejected, the result here is 2.420274 there seems to be no multicollinearity relationship in the model, so we can "transformation" the data of Year and Odometer at the same time, and finally find the right model.

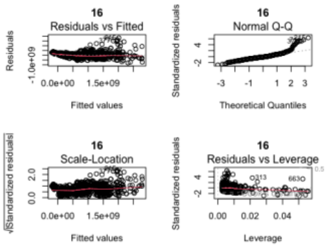
#f)

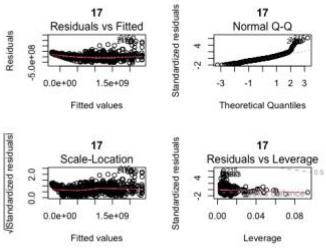
# Based on the information of lecture, if a polynomial term of a given order is retained, then all related terms of lower order are also retained in the model, and we don't really consider models with order higher than 3.

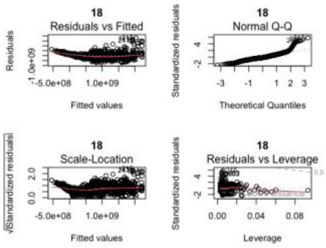


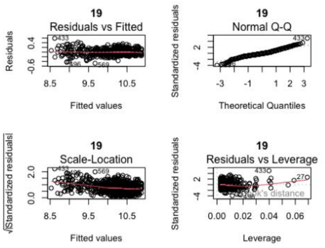


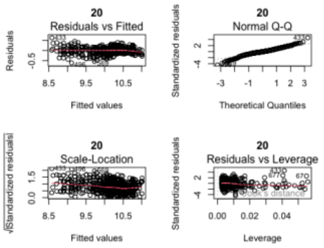


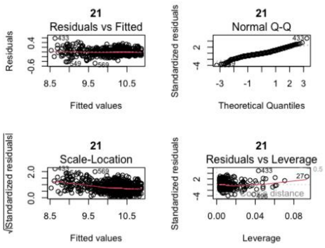


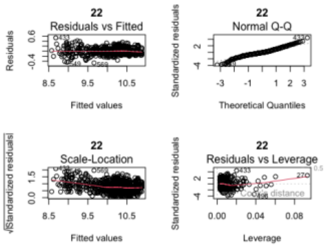












```
model1<-lm(AskPrice~ Year+ I(log(Year)) + Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model1,main = "1")

model1_1<-lm(AskPrice~ I(log(Year)) + Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model1_1,main = "1_1")

model2<-lm(AskPrice~ Year + Model + Odometer+ I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model2, main = "2")

model2_2<-lm(AskPrice~ Year + Model +  I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model2_2,main="2_2")

model3<-lm((log(AskPrice))~Year + Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model3, main="3")

model4<-lm(AskPrice~  I(log(Year)) + Model +  I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model4, main="4")

model4_2<-lm(AskPrice~ Year+ I(log(Year)) + Model + Odometer+ I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model4_2, main = "4_2")
```

```r
model5<-lm(AskPrice ~Year + I(Year^2) +Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model5, main= "5")

model6<-lm(AskPrice ~Year + I(Year^2)+ Model + I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model6 , main= "6")

model7<-lm(AskPrice ~Year + I(Year^2) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model7, main = "7")

model8<-lm(AskPrice ~I(log(Year)) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model8, main = "8")

model9<-lm((AskPrice^2)~Year + Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model9, main= "9")

model10<-lm(sqrt(AskPrice)~ Year + Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model10, main="10")

model11<-lm(sqrt(AskPrice)~ Year + I(Year^2) +Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model11, main="11")

model12<-lm(sqrt(AskPrice)~ Year + I(Year^2)+ Model + I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model12, main="12")

model13<-lm(sqrt(AskPrice)~ Year + I(Year^2) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model13, main="13")

model14<-lm(sqrt(AskPrice)~ (log(Year)) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model14, main="14")

model15<-lm((AskPrice^2)~ Year + I(Year^2) +Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model15, main="15")

model16<-lm((AskPrice^2)~ Year + I(Year^2)+ Model + I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model16, main="16")

model17<-lm((AskPrice^2)~ Year + I(Year^2) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model17, main="17")
```

```
model18<-lm((AskPrice^2)~ (log(Year)) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model18, main="18")

model19<-lm((log(AskPrice))~Year + I(Year^2) +Model + Odometer + Power)
par(mfrow=c(2,2))
plot(model19, main="19")

model20<-lm((log(AskPrice))~Year + I(Year^2)+ Model + I(log(Odometer)) + Power)
par(mfrow=c(2,2))
plot(model20, main="20")

model21<-lm((log(AskPrice))~Year + I(Year^2) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model21, main="21")

model22<-lm((log(AskPrice))~(log(Year)) +Model + Odometer+I(Odometer^2) + Power)
par(mfrow=c(2,2))
plot(model22, main="22")

#We used the "Residuals VS Fitted" image to select the more suitable candidates. Candidates selected according
to the chart, respectively mod: 1,3,4_2,5,7,11,12,13,14,19,20,21,22
summary(model1)

##
## Call:
## lm(formula = AskPrice ~ Year + I(log(Year)) + Model + Odometer +
##     Power)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9774.0 -2314.0  -554.6  1656.9 15680.5
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.085e+09  2.915e+08   14.02   <2e-16 ***
## Year           3.087e+05  2.192e+04   14.09   <2e-16 ***
## I(log(Year))  -6.187e+08  4.411e+07  -14.03   <2e-16 ***
## ModelCorolla  -4.234e+03  3.073e+02  -13.78   <2e-16 ***
## Odometer      -4.651e-02  3.454e-03  -13.46   <2e-16 ***
## PowerPetrol   -7.315e+03  2.930e+02  -24.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3556 on 812 degrees of freedom
## Multiple R-squared:  0.903,  Adjusted R-squared:  0.9024
## F-statistic:  1512 on 5 and 812 DF,  p-value: < 2.2e-16

summary(model3)

##
## Call:
```

```
## lm(formula = (log(AskPrice)) ~ Year + Model + Odometer + Power)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46337 -0.07326 -0.00904  0.06868  0.52664
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.402e+02  3.161e+00  -44.38   <2e-16 ***
## Year          7.475e-02  1.563e-03   47.82   <2e-16 ***
## ModelCorolla -1.457e-01  1.053e-02  -13.84   <2e-16 ***
## Odometer     -2.040e-06  1.166e-07  -17.49   <2e-16 ***
## PowerPetrol  -2.039e-01  9.901e-03  -20.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1222 on 813 degrees of freedom
## Multiple R-squared:  0.9377, Adjusted R-squared:  0.9373
## F-statistic:  3057 on 4 and 813 DF,  p-value: < 2.2e-16
```

```
summary(model4_2)
```

```
##
## Call:
## lm(formula = AskPrice ~ Year + I(log(Year)) + Model + Odometer +
##     I(log(Odometer)) + Power)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9886.8 -2241.0  -549.3  1668.0 15763.4
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.849e+09  3.201e+08  12.025   <2e-16 ***
## Year              2.910e+05  2.407e+04  12.088   <2e-16 ***
## I(log(Year))     -5.830e+08  4.845e+07 -12.033   <2e-16 ***
## ModelCorolla     -4.235e+03  3.069e+02 -13.799   <2e-16 ***
## Odometer         -4.353e-02  3.837e-03 -11.347   <2e-16 ***
## I(log(Odometer)) -2.261e+02  1.276e+02  -1.772   0.0768 .
## PowerPetrol      -7.281e+03  2.932e+02 -24.829   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3552 on 811 degrees of freedom
## Multiple R-squared:  0.9034, Adjusted R-squared:  0.9026
## F-statistic:  1263 on 6 and 811 DF,  p-value: < 2.2e-16
```

```
summary(model5)
```

```
## 
## Call:
## lm(formula = AskPrice ~ Year + I(Year^2) + Model + Odometer + 
##     Power)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9778.9 -2314.1  -553.5  1650.6 15677.5 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   3.067e+08  2.205e+07   13.91   <2e-16 ***
## Year         -3.060e+05  2.191e+04  -13.97   <2e-16 ***
## I(Year^2)     7.635e+01  5.442e+00   14.03   <2e-16 ***
## ModelCorolla -4.235e+03  3.073e+02  -13.78   <2e-16 ***
## Odometer     -4.652e-02  3.454e-03  -13.47   <2e-16 ***
## PowerPetrol  -7.314e+03  2.930e+02  -24.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3556 on 812 degrees of freedom
## Multiple R-squared:  0.903,  Adjusted R-squared:  0.9024 
## F-statistic:  1512 on 5 and 812 DF,  p-value: < 2.2e-16
```

```
summary(model7)
```

```
## 
## Call:
## lm(formula = AskPrice ~ Year + I(Year^2) + Model + Odometer + 
##     I(Odometer^2) + Power)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9789.4 -2268.4  -489.3  1635.8 15325.8 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    2.604e+08  2.445e+07  10.650  < 2e-16 ***
## Year          -2.600e+05  2.430e+04 -10.702  < 2e-16 ***
## I(Year^2)      6.492e+01  6.036e+00  10.755  < 2e-16 ***
## ModelCorolla  -4.259e+03  3.043e+02 -13.998  < 2e-16 ***
## Odometer      -7.839e-02  8.325e-03  -9.416  < 2e-16 ***
## I(Odometer^2)  1.299e-07  3.095e-08   4.199 2.98e-05 ***
## PowerPetrol   -7.243e+03  2.905e+02 -24.929  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3520 on 811 degrees of freedom
## Multiple R-squared:  0.9051, Adjusted R-squared:  0.9044 
## F-statistic:  1289 on 6 and 811 DF,  p-value: < 2.2e-16
```

```
summary(model11)
```

```
##
## Call:
## lm(formula = sqrt(AskPrice) ~ Year + I(Year^2) + Model + Odometer +
##     Power)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.491  -6.177  -1.229   4.436  35.131
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.284e+05  6.011e+04   7.127 2.27e-12 ***
## Year         -4.304e+02  5.974e+01  -7.205 1.33e-12 ***
## I(Year^2)     1.081e-01  1.484e-02   7.286 7.55e-13 ***
## ModelCorolla -1.220e+01  8.380e-01 -14.553  < 2e-16 ***
## Odometer     -1.537e-04  9.419e-06 -16.314  < 2e-16 ***
## PowerPetrol  -1.950e+01  7.988e-01 -24.409  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.697 on 812 degrees of freedom
## Multiple R-squared:  0.9275, Adjusted R-squared:  0.927
## F-statistic:  2077 on 5 and 812 DF,  p-value: < 2.2e-16
```

```
summary(model12)
```

```
##
## Call:
## lm(formula = sqrt(AskPrice) ~ Year + I(Year^2) + Model + I(log(Odometer)) +
##     Power)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.849  -6.476  -0.697   5.980  37.894
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.120e+05  7.458e+04   5.524 4.47e-08 ***
## Year             -4.153e+02  7.412e+01  -5.602 2.90e-08 ***
## I(Year^2)         1.047e-01  1.841e-02   5.683 1.84e-08 ***
## ModelCorolla     -1.122e+01  9.427e-01 -11.898  < 2e-16 ***
## I(log(Odometer)) -2.083e+00  3.537e-01  -5.891 5.62e-09 ***
## PowerPetrol      -1.985e+01  9.031e-01 -21.974  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.94 on 812 degrees of freedom
## Multiple R-squared:  0.9076,  Adjusted R-squared:  0.9071
## F-statistic:  1596 on 5 and 812 DF,  p-value: < 2.2e-16
```

```
summary(model13)
```

```
##
## Call:
## lm(formula = sqrt(AskPrice) ~ Year + I(Year^2) + Model + Odometer +
##     I(Odometer^2) + Power)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.503  -6.057  -1.124   4.522  35.014
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.727e+05  6.725e+04   5.542 4.04e-08 ***
## Year          -3.750e+02  6.683e+01  -5.612 2.75e-08 ***
## I(Year^2)      9.436e-02  1.660e-02   5.683 1.84e-08 ***
## ModelCorolla  -1.222e+01  8.369e-01 -14.607  < 2e-16 ***
## Odometer      -1.920e-04  2.290e-05  -8.387  < 2e-16 ***
## I(Odometer^2)  1.565e-10  8.513e-11   1.839   0.0663 .
## PowerPetrol   -1.941e+01  7.991e-01 -24.294  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.682 on 811 degrees of freedom
## Multiple R-squared:  0.9278,  Adjusted R-squared:  0.9272
## F-statistic:  1736 on 6 and 811 DF,  p-value: < 2.2e-16
```

```
summary(model14)
```

```
##
## Call:
## lm(formula = sqrt(AskPrice) ~ (log(Year)) + Model + Odometer +
##     I(Odometer^2) + Power)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.148  -6.211  -1.110   4.670  37.687
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -7.335e+04  1.940e+03 -37.815  < 2e-16 ***
## log(Year)      9.666e+03  2.548e+02  37.931  < 2e-16 ***
## ModelCorolla  -1.193e+01  8.517e-01 -14.005  < 2e-16 ***
## Odometer      -2.553e-04  2.048e-05 -12.471  < 2e-16 ***
## I(Odometer^2)  3.776e-10  7.746e-11   4.874 1.32e-06 ***
## PowerPetrol   -2.004e+01  8.072e-01 -24.824  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.872 on 812 degrees of freedom
## Multiple R-squared:  0.9248, Adjusted R-squared:  0.9244
## F-statistic:  1998 on 5 and 812 DF,  p-value: < 2.2e-16
```

```
summary(model19)
```

```
##
## Call:
## lm(formula = (log(AskPrice)) ~ Year + I(Year^2) + Model + Odometer +
##     Power)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47831 -0.07231 -0.01400  0.06262  0.59127
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.976e+03  7.516e+02  -3.960 8.16e-05 ***
## Year          2.893e+00  7.469e-01   3.873 0.000116 ***
## I(Year^2)    -7.001e-04  1.855e-04  -3.773 0.000173 ***
## ModelCorolla -1.426e-01  1.048e-02 -13.610  < 2e-16 ***
## Odometer     -2.123e-06  1.178e-07 -18.030  < 2e-16 ***
## PowerPetrol  -2.107e-01  9.988e-03 -21.099  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1212 on 812 degrees of freedom
## Multiple R-squared:  0.9387, Adjusted R-squared:  0.9384
## F-statistic:  2488 on 5 and 812 DF,  p-value: < 2.2e-16
```

```
summary(model20)
```

```
##
## Call:
## lm(formula = (log(AskPrice)) ~ Year + I(Year^2) + Model + I(log(Odometer)) +
##     Power)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61353 -0.07048 -0.00145  0.07802  0.62860
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.600e+03  9.637e+02  -2.698  0.00711 **
## Year              2.502e+00  9.577e-01   2.613  0.00915 **
## I(Year^2)        -5.988e-04  2.379e-04  -2.516  0.01205 *
## ModelCorolla     -1.283e-01  1.218e-02 -10.536  < 2e-16 ***
## I(log(Odometer)) -2.241e-02  4.570e-03  -4.904 1.13e-06 ***
## PowerPetrol      -2.169e-01  1.167e-02 -18.589  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1414 on 812 degrees of freedom
## Multiple R-squared:  0.9167, Adjusted R-squared:  0.9162
## F-statistic:  1786 on 5 and 812 DF,  p-value: < 2.2e-16
```

```
summary(model21)
```

```
##
## Call:
## lm(formula = (log(AskPrice)) ~ Year + I(Year^2) + Model + Odometer +
##     I(Odometer^2) + Power)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47820 -0.07101 -0.01267  0.06159  0.58242
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.506e+03  8.417e+02  -2.978  0.00299 **
## Year            2.426e+00  8.365e-01   2.900  0.00383 **
## I(Year^2)      -5.840e-04  2.078e-04  -2.810  0.00507 **
## ModelCorolla   -1.423e-01  1.048e-02 -13.588  < 2e-16 ***
## Odometer       -1.800e-06  2.866e-07  -6.279 5.54e-10 ***
## I(Odometer^2) -1.319e-12  1.066e-12  -1.238  0.21606
## PowerPetrol    -2.115e-01  1.000e-02 -21.142  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1212 on 811 degrees of freedom
## Multiple R-squared:  0.9388, Adjusted R-squared:  0.9384
## F-statistic:  2075 on 6 and 811 DF,  p-value: < 2.2e-16
```

```
summary(model22)
```

```
##
## Call:
## lm(formula = (log(AskPrice)) ~ (log(Year)) + Model + Odometer +
##     I(Odometer^2) + Power)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46847 -0.07116 -0.01324  0.06667  0.53213
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.139e+03  2.391e+01 -47.657  < 2e-16 ***
## log(Year)       1.511e+02  3.141e+00  48.112  < 2e-16 ***
## ModelCorolla   -1.441e-01   1.050e-02  -13.728   < 2e-16 ***
## Odometer        -1.425e-06   2.524e-07   -5.647  2.26e-08 ***
## I(Odometer^2)  -2.627e-12   9.547e-13   -2.752   0.00606 **
## PowerPetrol    -2.078e-01   9.948e-03  -20.885   < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1217 on 812 degrees of freedom
## Multiple R-squared:  0.9383,  Adjusted R-squared:  0.9379
## F-statistic:  2469 on 5 and 812 DF,  p-value: < 2.2e-16
```

```
#(i) Based on the value of "Adjusted R-squared" calculated by the summary() function, we pick the largest models, which should be model19(Adjusted R-squared=0.9384) and model21(Adjusted R-squared=0.9384).
# model19<-lm((log(AskPrice))~Year + I(Year^2) +Model + Odometer + Power)
# model21<-lm((log(AskPrice))~Year + I(Year^2) +Model + Odometer+I(Odometer^2) + Power)
# The MLR reflected by these two mods is shown above.
# (ii)
# The F-statistic is a statistic used to test whether the linear regression model is set appropriately. A larger F-statistic indicates a better overall fit of the regression model. And the F-statistic is calculated based on the ratio between the model's sum of squared residuals (SSE) and the model's sum of squared regressions (SSR).
# The p-value is the probability value associated with the F-statistic and is used to assess the significance of the F-statistic. Typically, we use the p-value for hypothesis testing. If the p-value is less than some pre-determined level of significance (usually 0.05), we reject the original hypothesis and conclude that the model is significant overall. In the case of MLR, the summary() function helps to obtain the results of the significance test for each independent variable, if the p-value is <0.05, there is reason to believe that the variables are not statistically significant, which can help to choose whether the model is reasonable or not.
# In summary, the F-statistic and the p-value together provide information on whether the overall goodness of fit of the linear regression model is significant. Both the larger F-statistic and the smaller p-value imply that the model as a whole is significant, i.e. that the combination of independent variables explains the dependent variable in a statistically significant way.

#(iii)
#Based on the data from the summary() function, it is known that:
# In model19, which measures the significance of Year and Year^2, their p-value is 0.000116, 0.000173 respectively, both less than 0.05 (the default significance level), and for Odometer the significance of p-value is < 2e-16, which is also insignificant.
# In model21, which measures the significance of Year and Year^2, their p-values are 0.003830.00507, respectively, with p-value > 0.05 (default significant level) for Year^2, which is considered significant as a variable in the MLR case, for significance of Odometer and Odometer^2, the p-values are 5.54e-10, 0.21606, respectively, with p-value>0.05 (default significant level) for Odometer^2, which is considered significant as a variable in the MLR case.
# Therefore, in the case where the adjusted R- squared of both models is a maximum of 0.9384, the model21 with the more significant variable should be chosen
```

#g)

```
# According to part f) we have chosen "model21<-lm((log(AskPrice))~Year + I(Year^2) + Model + Odometer + I(Odometer^2) + Power)". The data first needs to be acquired and analysed using the predict data, with the significant level using the default of 5%. In addition, since the dependent variable "AskPrice" is transformed by log() in the model, the prediction value should also be manipulated using the exp() function.
```

```r
test_data<-read.csv("test_data.csv")
Prediction_AskPirce <- predict(model21, newdata = data.frame(test_data), interval = "prediction", level = 0.95)
exp(Prediction_AskPirce)
```

```
##        fit      lwr      upr
## 1 16237.07 12785.70 20620.11
## 2 26714.30 21039.07 33920.40
## 3 31909.64 25144.09 40495.60
## 4 39430.24 31038.76 50090.40
## 5 26830.34 21142.60 34048.19
## 6 29291.05 23081.92 37170.45
## 7 44896.87 35359.71 57006.39
## 8 47345.11 37259.90 60160.10
```

```r
# The result generates a price range for AskPrice, given the mod21 we set, and the price range is the predicted value obtained based on the new data.
```