honor statement: "I have completed this work independently. The solutions given are entirely my own work."

1**. CARPRICE. Find on the D2L a car price dataset. Use R to perform a regression analysis on the dataset Your submission should take the form of a technical report and should consider the following:**

a. (10 pts.) Paste your final model into your submission (just the R output).

```
> CARPRICE$strok_horsepower<-CARPRICE$stroke*CARPRICE$horsepower
> CARPRICE<-CARPRICE%>%
+   select(carwidth,curbweight,stroke,compressionratio,horsepower,price)
> CARPRICE$horsepowerSQ<-(CARPRICE$horsepower)^2
> CARPRICE$curbweightSQ<-(CARPRICE$curbweight)^2
> CARPRICE$carwidthSQ<-(CARPRICE$carwidth)^2
> CARPRICE$strok_horsepower<-CARPRICE$stroke*CARPRICE$horsepower
> model_final <- lm(price~., data = CARPRICE)
> summary(model_final)

Call:
lm(formula = price ~ ., data = CARPRICE)

Residuals:
   Min      1Q  Median      3Q     Max
-8597.6 -1480.0  -248.7  1139.2 16397.1

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.488e+05  2.171e+05   3.450 0.000687 ***
carwidth       -2.281e+04  6.597e+03  -3.458 0.000668 ***
curbweight      -1.694e+01  5.702e+00  -2.971 0.003341 **
stroke          3.130e+03  2.722e+03   1.150 0.251605
compressionratio  2.329e+02  7.969e+01   2.922 0.003886 **
horsepower      3.626e+02  9.150e+01   3.963 0.000104 ***
horsepowerSQ    -3.803e-01  1.298e-01  -2.931 0.003785 **
curbweightSQ     4.013e-03  9.947e-04   4.034 7.86e-05 ***
carwidthSQ      1.736e+02  4.915e+01   3.531 0.000516 ***
strok_horsepower -4.781e+01  2.384e+01  -2.005 0.046322 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3278 on 195 degrees of freedom
```

Multiple R-squared: 0.8391, Adjusted R-squared: 0.8317
F-statistic: 113 on 9 and 195 DF, p-value: < 2.2e-16

b. (10 pts.) Describe the model building process through which you generated this model.

1.
Checking the summary of the dataset
2.
By checking the P-value of each variable to see which variables should be removed because their p-values are too large to reject the null hypothesis.
3.
After a sequence of try I would keep variable carwidth, curbweight, stroke ,compressionratio ,horsepower and price whose p-value is good enough to reject the null hypothesis and accept the alternative one which is H is not equal to zero
4.
Then I would check the correlation among variables , we could see what are the good predictors for the car price by checking the correlation table.
5.
Then I would check the plot to look for which variables should have second-order terms.
I would say horsepower, curbweight, and carwidth should have second-order term .
6.
By adding the second-term with variables -horsepower, curbweight, carwidth, AD R-squared has been improved from 78.3% to 82.9%
7.
Last step is that I would create an interaction term by picking the stroke and horsepower of these two variables based on my intuition. The AD R-squared has been improved from 82.9% to 83.1%, this is my final model.

c. (10 pts.) What significant second-order terms did you find, if any? Did you try all second-order terms? Did you look at scatter plots to determine which second-order terms to evaluate? Discuss the benefits and drawbacks of these two strategies.

By checking the plot , we could see which variables should have second-order terms.
Horsepower, curbweight, and carwidth these three variables has a linear , positive and strong relationship with the price, so I guess they might be good predictors for price and pick them as the second-term into the model.
By adding the second-term with variables -horsepower, curbweight, carwidth, AD R-squared has been improved from 78.3% to 82.9%.

d. (10 pts.) What significant interaction terms did you find, if any? Did you try all combinations of interaction terms? Do you think that is an appropriate strategy? What happens to the number of interaction terms as the number of independent terms increases?

Choosing which model to use will depend in part on our domain knowledge and in part on the regression analysis.
Based on my intuition, I would try the interaction terms by picking the variable stroke and horsepower since they are kind of the car devices .

The AD R-squared has been improved from 82.9% to 83.1%.

e. (10 pts.) Discuss your final model. Evaluate the t-tests, F-Test and adj-R2 accordingly. Do you think this is a "good" model? Explain.

I would say it is a good model ,because it has a good adj-R2 which is 83.17%.

F-Test is good, reject the null hypothesis , accept the alternative β which is at least one β not equal to zero.

adj-R2 is 0.8317 which means that 83.17% of variability in price is explained by the model.

All the T-test look good except for the t-test of stroke, but the t-test of interaction term is good and we should include the child of the interaction term, so I would not exclude the stroke variable.

I would say it is a good model ,because it has a good adj-R2 which is 83.17%.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.488e+05	2.171e+05	3.450	0.000687 ***
carwidth	-2.281e+04	6.597e+03	-3.458	0.000668 ***
curbweight	-1.694e+01	5.702e+00	-2.971	0.003341 **
stroke	3.130e+03	2.722e+03	1.150	0.251605
compressionratio	2.329e+02	7.969e+01	2.922	0.003886 **
horsepower	3.626e+02	9.150e+01	3.963	0.000104 ***
horsepowerSQ	-3.803e-01	1.298e-01	-2.931	0.003785 **
curbweightSQ	4.013e-03	9.947e-04	4.034	7.86e-05 ***
carwidthSQ	1.736e+02	4.915e+01	3.531	0.000516 ***
strok_horsepower	-4.781e+01	2.384e+01	-2.005	0.046322 *

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3278 on 195 degrees of freedom
Multiple R-squared:  0.8391,  Adjusted R-squared:  0.8317
F-statistic:   113 on 9 and 195 DF,  p-value: < 2.2e-16
```

f. Include your code an appendix.

```
> CARPRICE<-CARPRICE%>%
+
select(wheelbase,carlength,carwidth,carheight,curbweight,boreratio,stroke,compressionratio,hor
sepower,peakrpm,citympg,highwaympg,price)
> model <- lm(price~., data = CARPRICE)
> summary(model)

Call:
lm(formula = price ~ ., data = CARPRICE)

Residuals:
   Min      1Q  Median      3Q     Max
-9053.7 -1891.3  -122.9  1533.2 15769.3

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.432e+04  1.786e+04  -2.481 0.013965 *
wheelbase       1.472e+02  1.175e+02   1.252 0.211950
carlength      -7.748e+01  6.497e+01  -1.193 0.234491
carwidth        5.097e+02  2.879e+02   1.770 0.078244 .
carheight      -9.467e+00  1.570e+02  -0.060 0.951989
curbweight      6.774e+00  1.918e+00   3.532 0.000517 ***
boreratio      -1.289e+03  1.399e+03  -0.921 0.357966
stroke         -1.955e+03  8.989e+02  -2.175 0.030830 *
compressionratio  1.573e+02  9.506e+01   1.655 0.099556 .
horsepower      1.041e+02  1.605e+01   6.486 7.31e-10 ***
peakrpm         3.596e-01  7.341e-01   0.490 0.624779
citympg         6.971e+00  2.031e+02   0.034 0.972651
highwaympg      7.644e+01  1.861e+02   0.411 0.681766
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3732 on 192 degrees of freedom
```

Multiple R-squared: 0.7946, Adjusted R-squared: 0.7818
F-statistic: 61.9 on 12 and 192 DF, p-value: < 2.2e-16

By checking the P-value of each variable, I would remove citympg , because its p-value is too large to reject the null hypothesis.

```
> CARPRICE<-CARPRICE%>%
+
select(wheelbase,carlength,carwidth,carheight,curbweight,boreratio,stroke,compressionratio,hor
sepower,peakrpm,highwaympg,price)
> model2 <- lm(price~., data = CARPRICE)
> summary(model2)

Call:
lm(formula = price ~ ., data = CARPRICE)

Residuals:
    Min      1Q  Median      3Q     Max
-9070.5 -1878.9  -130.6  1532.5 15771.4


Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.426e+04  1.774e+04  -2.495 0.013449 *
wheelbase       1.482e+02  1.135e+02   1.306 0.193253
carlength      -7.815e+01  6.184e+01  -1.264 0.207833
carwidth        5.094e+02  2.870e+02   1.775 0.077518 .
carheight      -9.369e+00  1.566e+02  -0.060 0.952354
curbweight      6.780e+00  1.904e+00   3.560 0.000467 ***
boreratio      -1.293e+03  1.389e+03  -0.931 0.352967
stroke         -1.958e+03  8.937e+02  -2.191 0.029665 *
compressionratio  1.580e+02  9.294e+01   1.700 0.090803 .
horsepower      1.040e+02  1.560e+01   6.665  2.7e-10 ***
peakrpm         3.575e-01  7.295e-01   0.490 0.624682
highwaympg      8.207e+01  8.783e+01   0.934 0.351245
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3722 on 193 degrees of freedom
Multiple R-squared:  0.7946,  Adjusted R-squared:  0.7829
```

F-statistic: 67.88 on 11 and 193 DF, p-value: < 2.2e-16

By checking the P-value of each variable in model 2, I would remove carheight , because its p-value is too large to reject the null hypothesis.

```
> CARPRICE<-CARPRICE%>%
+
select(wheelbase,carlength,carwidth,curbweight,boreratio,stroke,compressionratio,horsepower,
peakrpm,highwaympg,price)
> model3 <- lm(price~., data = CARPRICE)
> summary(model3)

Call:
lm(formula = price ~ ., data = CARPRICE)

Residuals:
    Min      1Q  Median      3Q     Max
-9092.9 -1873.5  -121.1  1532.3 15776.7

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.467e+04  1.632e+04  -2.736 0.006790 **
wheelbase       1.459e+02  1.068e+02   1.367 0.173307
carlength      -7.909e+01  5.967e+01  -1.326 0.186551
carwidth        5.133e+02  2.787e+02   1.842 0.067022 .
curbweight      6.777e+00  1.899e+00   3.569 0.000452 ***
boreratio      -1.291e+03  1.385e+03  -0.932 0.352454
stroke         -1.948e+03  8.754e+02  -2.225 0.027235 *
compressionratio  1.578e+02  9.265e+01   1.703 0.090169 .
horsepower      1.042e+02  1.533e+01   6.793 1.31e-10 ***
peakrpm         3.568e-01  7.275e-01   0.490 0.624379
highwaympg      8.166e+01  8.734e+01   0.935 0.350940
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3713 on 194 degrees of freedom
Multiple R-squared:  0.7946,  Adjusted R-squared:  0.784
F-statistic: 75.05 on 10 and 194 DF,  p-value: < 2.2e-16
```

By checking the P-value of each variable in model 3, I would remove peakrpm , because its p-value is too large to reject the null hypothesis.

After a sequence of try I would keep variable carwidth, curbweight, stroke ,compressionratio ,horsepower and price whose p-vlaue is good enough to reject the null hypothesis and accept the alternative one which is H is not equal to zero

```
> CARPRICE<-CARPRICE%>%
+   select(carwidth,curbweight,stroke,compressionratio,horsepower,price)
> modelfinal <- lm(price~., data = CARPRICE)
> summary(modelfinal)

Call:
lm(formula = price ~ ., data = CARPRICE)

Residuals:
   Min      1Q  Median      3Q     Max
-7859.8 -2228.2  -112.9  1444.1 15687.1

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -43836.031  14025.751  -3.125  0.00204 **
carwidth           559.579    245.219   2.282  0.02355 *
curbweight           5.158      1.209   4.265 3.08e-05 ***
stroke           -1599.712    855.158  -1.871  0.06286 .
compressionratio   202.541     77.061   2.628  0.00925 **
horsepower          98.013     11.409   8.591 2.51e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3717 on 199 degrees of freedom
Multiple R-squared:  0.7888,  Adjusted R-squared:  0.7835
F-statistic: 148.7 on 5 and 199 DF,  p-value: < 2.2e-16
```

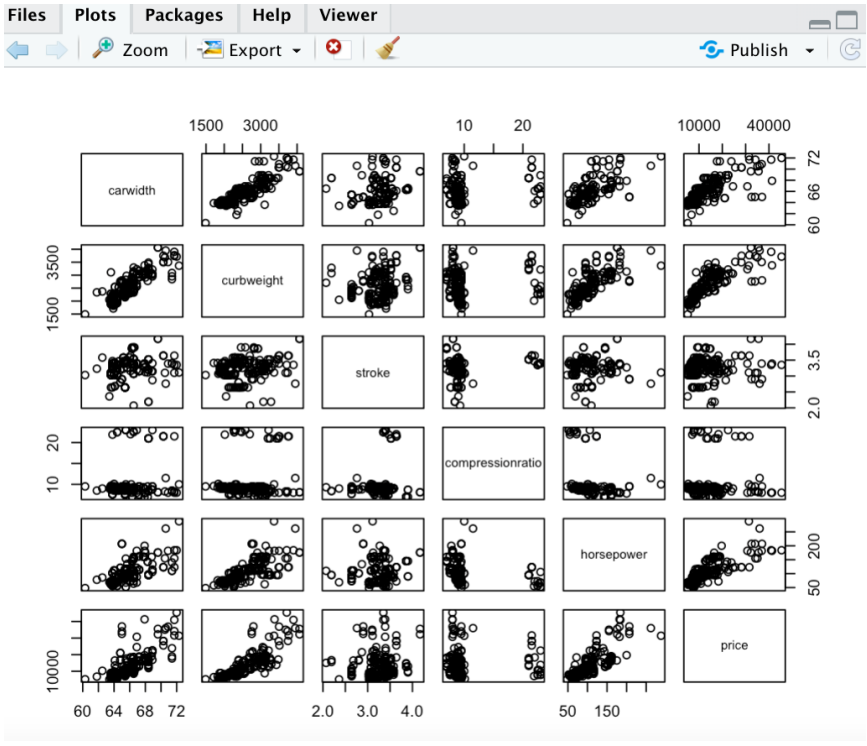
Then I would check the correlation among variable , we could see what are the good predictors for the car price by checking the correlation table.

```
> cor(CARPRICE)
                carwidth curbweight     stroke compressionratio  horsepower      price
carwidth        1.0000000  0.8670325 0.18294169       0.18112863  0.64073208 0.75932530
curbweight      0.8670325  1.0000000 0.16879004       0.15136174  0.75073925 0.83530488
stroke          0.1829417  0.1687900 1.00000000       0.18611011  0.08093954 0.07944308
compressionratio 0.1811286  0.1513617 0.18611011       1.00000000 -0.20432623 0.06798351
horsepower      0.6407321  0.7507393 0.08093954      -0.20432623  1.00000000 0.80813882
price           0.7593253  0.8353049 0.07944308       0.06798351  0.80813882 1.00000000
```

Then I would check the plot to look for which variables should have second-order terms.

I would say horsepower, curbweight, and carwidth should have second-order term .

```
plot(CARPRICE)
```



By adding the second-term with variables -horsepower, curbweight, carwidth, AD R-squared has been improved from 78.3% to 82.9%

```
> CARPRICE<-CARPRICE%>%
+   select(carwidth,curbweight,stroke,compressionratio,horsepower,price)
```

```
> CARPRICE$horsepowerSQ<-(CARPRICE$horsepower)^2
> CARPRICE$curbweightSQ<-(CARPRICE$curbweight)^2
> CARPRICE$carwidthSQ<-(CARPRICE$carwidth)^2
> modelfinal <- lm(price~., data = CARPRICE)
> summary(modelfinal)

Call:
lm(formula = price ~ ., data = CARPRICE)

Residuals:
   Min      1Q  Median      3Q     Max
-7232.0 -1611.3   -98.2  1178.0 15961.2

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.694e+05  2.185e+05   3.522 0.000534 ***
carwidth        -2.306e+04  6.647e+03  -3.470 0.000640 ***
curbweight       -1.445e+01  5.608e+00  -2.577 0.010715 *
stroke          -2.093e+03  7.970e+02  -2.626 0.009319 **
compressionratio  2.800e+02  7.673e+01   3.649 0.000337 ***
horsepower       1.978e+02  4.050e+01   4.883 2.17e-06 ***
horsepowerSQ     -3.156e-01  1.267e-01  -2.492 0.013533 *
curbweightSQ      3.434e-03  9.593e-04   3.580 0.000433 ***
carwidthSQ       1.759e+02  4.951e+01   3.552 0.000478 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3303 on 196 degrees of freedom
Multiple R-squared:  0.8358,  Adjusted R-squared:  0.8291
F-statistic: 124.7 on 8 and 196 DF,  p-value: < 2.2e-16
```

Last step is that I would create an interaction term by picking the stroke and horsepower of these two variables based on my intuition, the AD R-squared has been improved from 82.9% to 83.1%, this is my final model

```
> CARPRICE$strok_horsepower<-CARPRICE$stroke*CARPRICE$horsepower
> CARPRICE<-CARPRICE%>%
+   select(carwidth,curbweight,stroke,compressionratio,horsepower,price)
> CARPRICE$horsepowerSQ<-(CARPRICE$horsepower)^2
> CARPRICE$curbweightSQ<-(CARPRICE$curbweight)^2
> CARPRICE$carwidthSQ<-(CARPRICE$carwidth)^2
> CARPRICE$strok_horsepower<-CARPRICE$stroke*CARPRICE$horsepower
```

```
> model_final <- lm(price~., data = CARPRICE)
> summary(model_final)

Call:
lm(formula = price ~ ., data = CARPRICE)

Residuals:
   Min      1Q  Median      3Q     Max
-8597.6 -1480.0  -248.7  1139.2 16397.1

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.488e+05  2.171e+05   3.450 0.000687 ***
carwidth        -2.281e+04  6.597e+03  -3.458 0.000668 ***
curbweight       -1.694e+01  5.702e+00  -2.971 0.003341 **
stroke           3.130e+03  2.722e+03   1.150 0.251605
compressionratio  2.329e+02  7.969e+01   2.922 0.003886 **
horsepower       3.626e+02  9.150e+01   3.963 0.000104 ***
horsepowerSQ     -3.803e-01  1.298e-01  -2.931 0.003785 **
curbweightSQ      4.013e-03  9.947e-04   4.034 7.86e-05 ***
carwidthSQ       1.736e+02  4.915e+01   3.531 0.000516 ***
strok_horsepower -4.781e+01  2.384e+01  -2.005 0.046322 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3278 on 195 degrees of freedom
Multiple R-squared:  0.8391,  Adjusted R-squared:  0.8317
F-statistic:   113 on 9 and 195 DF,  p-value: < 2.2e-16
```