Toyota Corolla prediction

Xueyang Wu,Zhou zhou, Xinyu Wang, Ke Wu
08/05/2020

Abstract

Sale price for a car can be driven by several factors or characters in the car. This report analyzed the data on price of car sold by Toyota to see how much the price of car will differ under different characteristics. The variables we used to predict include The sales price of the Toyota Corolla (in Euros), Age of the purchaser, Odometer reading in kilometers, Fuel type (Diesel, Petrol or CNG), Horsepower, Color, Transmission (manual or automatic), Displacement in cubic centimeters, Number of doors, and weight in kilograms. The primary objective of this report is to find the drivers of sales price of cars. In this analysis, we found there was no obviously different distributions for any dummy variables. Through a regression analysis, it was found that the primary drivers of sales price for car was **Age, Weight, HP, KM, and CC**. In fact, we also found for buyers, one year older the buyers are, the lower 170.934 Euros they would pay for the car. In our report, it is also shown that if the car is automatic but not manual, it will be 522.93 Euros higher for the price.

Introduction

The dealership runs the business of selling cars directly to the customers, so they want to setup the suitable price that can drives customers to make a deal with them. If the price of cars with vary characteristics is not set at the appropriate level, dealership would hard to convince customer to make a purchase. Therefore, it's necessary to examine and evaluate how different characteristics in a car would influence the sales price for a car. This will help dealerships to make decision on set the proper price for cars with specific characteristics that purchasers will glad to pay but at the same time can still profit dealerships. In the dataset we researched on, not only characteristics for cars will influence the price, the characteristic of purchasers themselves (Age) is also taking into account. Through the data analysis we did, it's possible to observe the significance of the available variables and correlation of them with price, and to determine which variable(s) is(are) the primary drivers for sales price.

The data set consists of 9 independent variables including:

- Age – Age of the purchaser
- KM – Odometer reading in kilometers
- FuelType – Fuel type (Diesel, Petrol or CNG)
- HP – Horsepower
- MetColor – Color
- Automatic – Transmission (0 = Manual, 1 = Automatic)
- CC – Displacement in cubic centimeters

• Doors – Number of doors
• Weight – Weight in kilograms

Before we are starting the examination by using the scatterplot matrix, to decide which columns to use in matrix and cleanup the scatterplot shown up, we first started with the correlation plot to observe which variables have high correlation with Price.

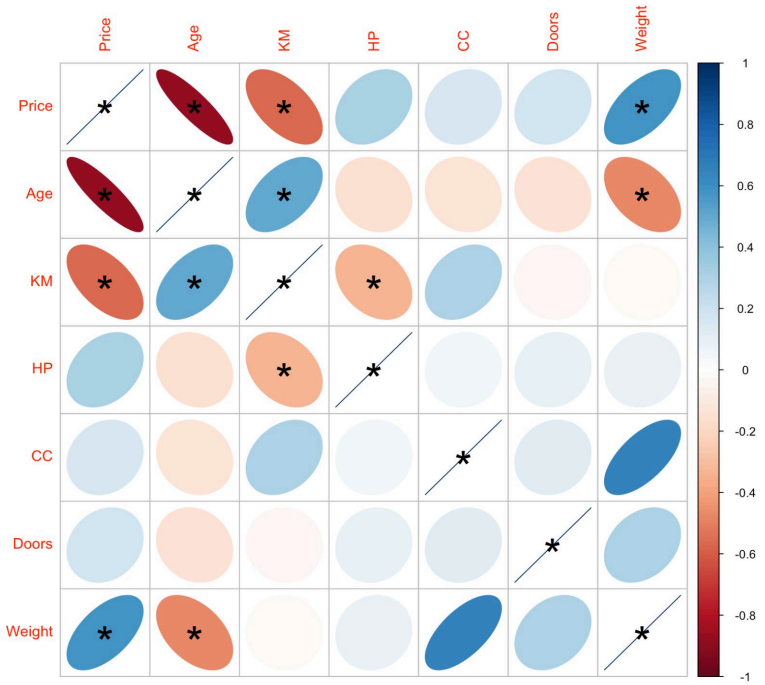


Figure1

In Figure1, we found that both Age and KM shown up a negative high correlation with Price, and Weight and HP shown up a relatively high positive correlation with Price. Based on this we decided to use these 4 variables to create a scatterplot matrix by using ggpair.

In the Figure 2,3,and 4(See Appendix) we colored the variables by dummy variables in dataset. (FuelType, MetColor, and Automastics) We didn't find anything special here, the graph for all three of them in Price*Price matrix shown up that they are highly coincide with other. These scatterplots indicate that sales price will not be different depends on the dummy variables, so we don't need to consider any dummy variables in separate models. Then we decided to start the analysis with the simple linear regression.

Analysis

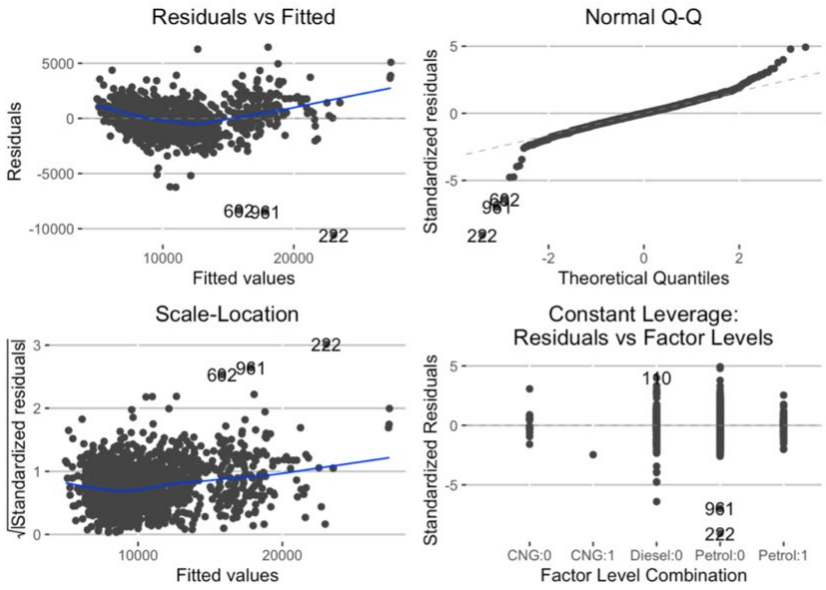
After analyzation we did before, we decided to start by doing a simple linear regression for all variables to see the significance of overall model first. If the model has the really low adjusted R-square, we need to try other method to improve the significance.

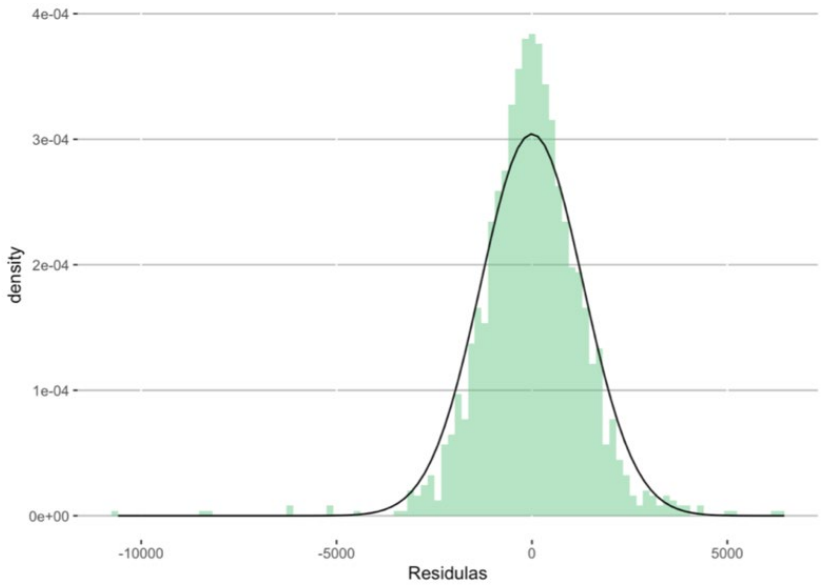
```
##
## Call:
## lm(formula = Price ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10642.3   -737.7      3.1    731.3   6451.5
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.801e+03  1.304e+03  -2.915 0.003613 **
## Age            -1.220e+02  2.602e+00 -46.889  < 2e-16 ***
## KM             -1.621e-02  1.313e-03 -12.347  < 2e-16 ***
## FuelTypeDiesel  3.390e+03  5.188e+02   6.535 8.86e-11 ***
## FuelTypePetrol  1.121e+03  3.324e+02   3.372 0.000767 ***
## HP              6.081e+01  5.756e+00  10.565  < 2e-16 ***
## MetColor1       5.716e+01  7.494e+01   0.763 0.445738
## Automatic1      3.303e+02  1.571e+02   2.102 0.035708 *
## CC             -4.174e+00  5.453e-01  -7.656 3.53e-14 ***
## Doors          -7.776e+00  4.006e+01  -0.194 0.846129
## Weight          2.001e+01  1.203e+00  16.629  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1316 on 1425 degrees of freedom
## Multiple R-squared:  0.8693, Adjusted R-squared:  0.8684
## F-statistic:   948 on 10 and 1425 DF,  p-value: < 2.2e-16
```

The initial Adjusted R-squared is 0.8684. The result we get is already relatively good enough, but there are some variables are insignificant. So, we decide to see the result after subtracting those variables.

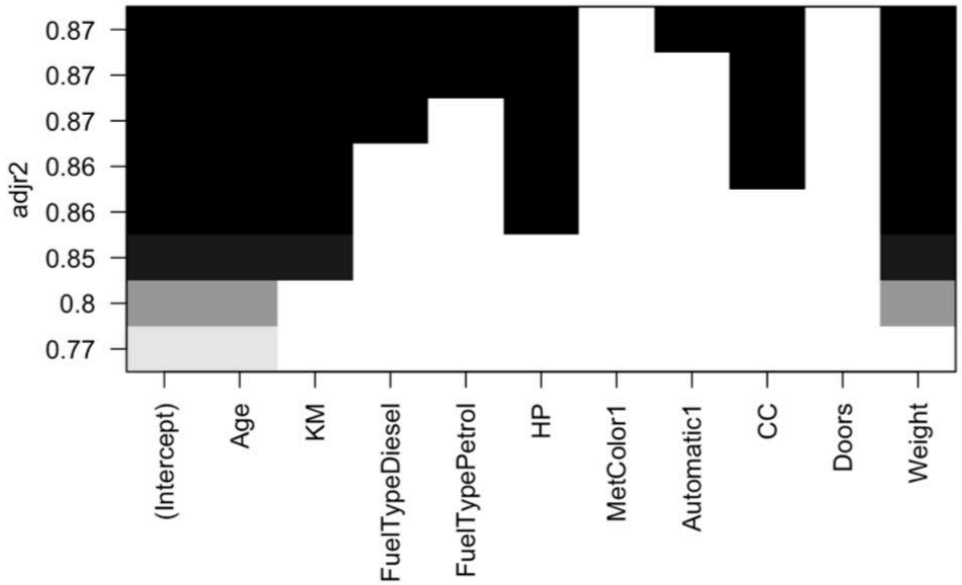
```
##
## Call:
## lm(formula = Price ~ . - Doors - MetColor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10593.9   -726.9     -2.3    720.1   6459.4
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.718e+03  1.261e+03  -2.948  0.00325 **
## Age            -1.221e+02  2.596e+00 -47.041  < 2e-16 ***
## KM             -1.625e-02  1.309e-03 -12.416  < 2e-16 ***
## FuelTypeDiesel  3.388e+03  5.090e+02   6.655 4.03e-11 ***
## FuelTypePetrol  1.112e+03  3.317e+02   3.353  0.00082 ***
## HP              6.089e+01  5.639e+00  10.799  < 2e-16 ***
## Automatic1      3.305e+02  1.562e+02   2.116  0.03452 *
## CC             -4.168e+00  5.369e-01  -7.763 1.57e-14 ***
## Weight          1.994e+01  1.126e+00  17.709  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1315 on 1427 degrees of freedom
## Multiple R-squared:  0.8693, Adjusted R-squared:  0.8685
## F-statistic:  1186 on 8 and 1427 DF,  p-value: < 2.2e-16
```

This is model 1. We delete the “Doors” and “MetColor” variables. The adjusted R-squared increases by 0.0001.





From the Residual vs Fitted plot and Q-Q plot, we can see most points formed an obvious trend. The data fit the model pretty well. But there still are some outliers that we need to deal with. So, we want to use Best Subset to see whether we can improve the model by changing the number of variables in the model.



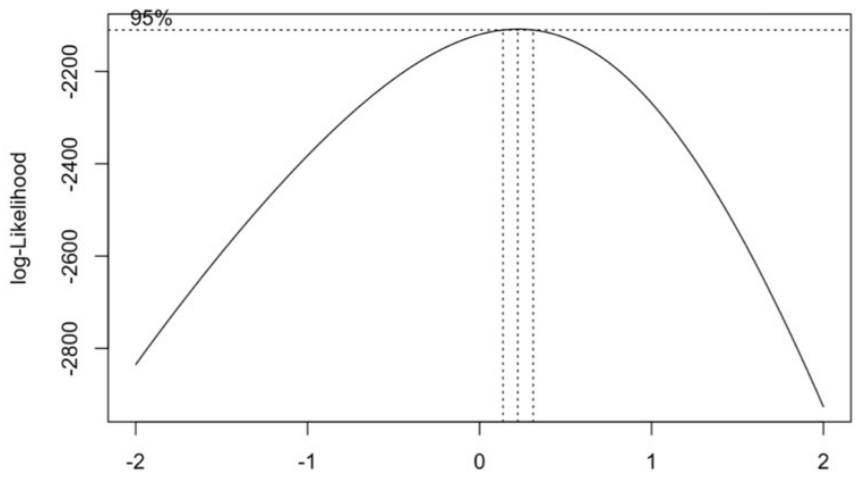

```
coef(BSR,8)
```

```
##    (Intercept)             Age              KM FuelTypeDiesel FuelTypePetrol
##  -3.718364e+03  -1.221299e+02  -1.625505e-02   3.387675e+03   1.112162e+03
##             HP      Automatic1              CC         Weight
##   6.089325e+01   3.304641e+02  -4.168202e+00   1.993833e+01
```

So, it turns out the best subset model is almost not different from our model 1. Then we use stepwise to verify whether variable number in our model is the best.

```
##
## Call:
## lm(formula = Price ~ Age + Weight + KM + HP + CC + FuelType +
##     Automatic, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10593.9   -726.9     -2.3    720.1   6459.4
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.718e+03  1.261e+03  -2.948  0.00325 **
## Age            -1.221e+02  2.596e+00 -47.041  < 2e-16 ***
## Weight          1.994e+01  1.126e+00  17.709  < 2e-16 ***
## KM             -1.625e-02  1.309e-03 -12.416  < 2e-16 ***
## HP              6.089e+01  5.639e+00  10.799  < 2e-16 ***
## CC             -4.168e+00  5.369e-01  -7.763 1.57e-14 ***
## FuelTypeDiesel  3.388e+03  5.090e+02   6.655 4.03e-11 ***
## FuelTypePetrol  1.112e+03  3.317e+02   3.353  0.00082 ***
## Automatic1      3.305e+02  1.562e+02   2.116  0.03452 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1315 on 1427 degrees of freedom
## Multiple R-squared:  0.8693, Adjusted R-squared:  0.8685
## F-statistic:  1186 on 8 and 1427 DF,  p-value: < 2.2e-16
```

This is model 3. The result of stepwise shown that there's still nothing changes from our model 1. So, we use Boxcox to see does the model need a transformation.



The lambda value is near to 0. So, we try to do a log(y) on the linear regression model.

```
##
## Call:
## lm(formula = log(Price) ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79015 -0.06463  0.00371  0.07351  0.46657
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.560e+00  1.148e-01  74.587  < 2e-16 ***
## Age            -1.053e-02  2.290e-04 -45.976  < 2e-16 ***
## KM             -1.694e-06  1.155e-07 -14.661  < 2e-16 ***
## FuelTypeDiesel  1.021e-01  4.565e-02   2.235  0.02555 *
## FuelTypePetrol  7.383e-02  2.925e-02   2.524  0.01170 *
## HP              2.800e-03  5.065e-04   5.528 3.84e-08 ***
## MetColor1       2.890e-03  6.595e-03   0.438  0.66127
## Automatic1      4.116e-02  1.382e-02   2.977  0.00296 **
## CC             -6.342e-05  4.798e-05  -1.322  0.18647
## Doors           6.866e-03  3.526e-03   1.947  0.05169 .
## Weight          1.013e-03  1.059e-04   9.566  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1158 on 1425 degrees of freedom
## Multiple R-squared:  0.8483, Adjusted R-squared:  0.8472
## F-statistic: 796.8 on 10 and 1425 DF,  p-value: < 2.2e-16
```

This is model 4. In this one we try to use log (Price) and test it with all other variables. But it seems like R-squared becomes smaller. So, we decide not to use log(y). Then we generate the idea of using the 2 interactions between all variables to create another model and see whether the R square result could be better.

Model 5:

```
##
## Call:
## lm(formula = Price ~ .^2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7579.4  -675.6   -26.8   683.2  6814.8
##
## Coefficients: (2 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.020e+06  1.458e+06  -0.700  0.48414
## Age                         7.505e+02  1.101e+02   6.814 1.41e-11 ***
## KM                         -9.748e-02  5.115e-02  -1.906  0.05690 .
## FuelTypeDiesel              9.717e+05  1.458e+06   0.667  0.50517
## FuelTypePetrol              9.930e+05  1.458e+06   0.681  0.49583
## HP                         -2.354e+02  2.093e+02  -1.125  0.26092
## MetColor1                   8.379e+03  3.056e+03   2.742  0.00618 **
## Automatic1                 -1.158e+04  6.633e+03  -1.746  0.08096 .
## CC                          6.490e+02  9.138e+02   0.710  0.47768
## Doors                       1.874e+03  1.889e+03   0.992  0.32126
## Weight                      2.401e+00  3.230e+01   0.074  0.94075
## Age:KM                      4.038e-04  6.959e-05   5.802 8.12e-09 ***
## Age:FuelTypeDiesel          1.574e+00  6.054e+01   0.026  0.97926
## Age:FuelTypePetrol         -4.265e+01  3.466e+01  -1.230  0.21872
## Age:HP                     -2.178e-02  7.895e-01  -0.028  0.97799
## Age:MetColor1              -1.292e+01  5.083e+00  -2.542  0.01112 *
## Age:Automatic1              2.025e+01  1.381e+01   1.467  0.14273
## Age:CC                     -2.804e-02  6.703e-02  -0.418  0.67581
## Age:Doors                   7.471e+00  2.720e+00   2.747  0.00609 **
## Age:Weight                 -7.746e-01  8.924e-02  -8.680  < 2e-16 ***
## KM:FuelTypeDiesel           5.419e-02  2.330e-02   2.326  0.02016 *
## KM:FuelTypePetrol           2.462e-02  1.041e-02   2.365  0.01816 *
## KM:HP                       4.875e-04  3.180e-04   1.533  0.12558
## KM:MetColor1                3.299e-03  2.422e-03   1.362  0.17350
## KM:Automatic1               1.676e-03  5.838e-03   0.287  0.77412
## KM:CC                      -4.221e-05  2.476e-05  -1.705  0.08837 .
## KM:Doors                   -1.314e-03  1.384e-03  -0.949  0.34271
## KM:Weight                   4.725e-05  5.237e-05   0.902  0.36703
## FuelTypeDiesel:HP          -5.991e+01  2.328e+01  -2.573  0.01018 *
## FuelTypePetrol:HP                  NA         NA      NA       NA
## FuelTypeDiesel:MetColor1    3.822e+03  1.306e+03   2.926  0.00349 **
## FuelTypePetrol:MetColor1    3.347e+02  1.030e+03   0.325  0.74528
## FuelTypeDiesel:Automatic1          NA         NA      NA       NA
## FuelTypePetrol:Automatic1   4.130e+03  1.591e+03   2.596  0.00954 **
## FuelTypeDiesel:CC          -6.321e+02  9.134e+02  -0.692  0.48907
```

```
## FuelTypePetrol:CC          -6.449e+02  9.134e+02  -0.706  0.48031
## FuelTypeDiesel:Doors        4.491e+02  7.802e+02   0.576  0.56494
## FuelTypePetrol:Doors       -1.892e+02  6.044e+02  -0.313  0.75433
## FuelTypeDiesel:Weight       2.800e+01  2.782e+01   1.006  0.31435
## FuelTypePetrol:Weight       3.570e+01  2.540e+01   1.406  0.16001
## HP:MetColor1                4.844e+01  1.213e+01   3.993 6.87e-05 ***
## HP:Automatic1              -8.435e+01  1.412e+02  -0.597  0.55043
## HP:CC                       2.869e-02  5.643e-02   0.508  0.61129
## HP:Doors                    3.722e-01  7.161e+00   0.052  0.95855
## HP:Weight                   1.393e-01  1.855e-01   0.751  0.45282
## MetColor1:Automatic1        1.510e+02  2.979e+02   0.507  0.61241
## MetColor1:CC               -2.890e+00  1.056e+00  -2.737  0.00628 **
## MetColor1:Doors             4.675e+01  7.925e+01   0.590  0.55539
## MetColor1:Weight           -8.543e+00  2.866e+00  -2.980  0.00293 **
## Automatic1:CC               3.969e+00  1.085e+01   0.366  0.71460
## Automatic1:Doors           -1.404e+02  1.795e+02  -0.782  0.43439
## Automatic1:Weight           8.930e+00  6.597e+00   1.354  0.17605
## CC:Doors                   -7.020e-01  6.228e-01  -1.127  0.25988
## CC:Weight                   3.199e-03  1.953e-02   0.164  0.86993
## Doors:Weight               -8.766e-01  1.644e+00  -0.533  0.59390
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1119 on 1383 degrees of freedom
## Multiple R-squared:  0.9083, Adjusted R-squared:  0.9048
## F-statistic: 263.3 on 52 and 1383 DF,  p-value: < 2.2e-16
```

This is model 5. We got a higher Adjusted R-squared value which is 0.9048. However, most of the variables are insignificant and we definitely want to improve this. Before we tried to improve the model by eliminating variables that are insignificant, we first try the square of some numerical variables to see whether the adjusted R square will went up more.

Model 6:

```
##
## Call:
## lm(formula = Price ~ .^2 + I(Age^2) + I(KM^2) + I(HP^2) + I(CC^2) +
##     I(Weight^2) + I(Doors^2) - Doors, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4832.4  -644.4   -29.5   646.3  6188.6
##
## Coefficients: (2 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.116e+06  1.400e+06  -0.797 0.425383
## Age                         7.769e+02  1.202e+02   6.461 1.44e-10 ***
## KM                         -1.283e-01  5.215e-02  -2.459 0.014040 *
## FuelTypeDiesel              8.200e+05  1.400e+06   0.586 0.558050
## FuelTypePetrol              9.526e+05  1.399e+06   0.681 0.496092
## HP                         -1.344e+03  4.313e+02  -3.116 0.001868 **
## MetColor1                   5.990e+03  2.949e+03   2.031 0.042445 *
## Automatic1                 -3.918e+03  6.428e+03  -0.610 0.542274
## CC                          7.887e+02  8.790e+02   0.897 0.369702
## Weight                      1.264e+02  3.230e+01   3.914 9.52e-05 ***
## I(Age^2)                   -5.199e-02  1.501e-01  -0.346 0.729058
## I(KM^2)                    -7.623e-08  2.569e-08  -2.967 0.003060 **
## I(HP^2)                    -1.220e+00  8.961e-01  -1.361 0.173682
## I(CC^2)                    -4.221e-02  2.567e-02  -1.645 0.100258
## I(Weight^2)                -5.957e-02  5.889e-03 -10.116  < 2e-16 ***
## I(Doors^2)                 -2.978e+02  1.092e+02  -2.727 0.006478 **
## Age:KM                      4.552e-04  9.727e-05   4.680 3.15e-06 ***
## Age:FuelTypeDiesel          9.655e+01  5.957e+01   1.621 0.105321
## Age:FuelTypePetrol         -3.355e+01  3.358e+01  -0.999 0.317921
## Age:HP                      9.656e-01  7.979e-01   1.210 0.226436
## Age:MetColor1              -1.182e+01  4.927e+00  -2.398 0.016616 *
## Age:Automatic1              1.183e+01  1.348e+01   0.877 0.380370
## Age:CC                     -1.828e-01  6.730e-02  -2.717 0.006673 **
## Age:Doors                   8.267e+00  2.443e+00   3.384 0.000733 ***
## Age:Weight                 -6.763e-01  9.957e-02  -6.792 1.64e-11 ***
## KM:FuelTypeDiesel           2.413e-02  2.304e-02   1.047 0.295268
## KM:FuelTypePetrol           1.884e-02  1.050e-02   1.795 0.072846 .
## KM:HP                       2.421e-05  3.192e-04   0.076 0.939541
## KM:MetColor1                2.633e-03  2.353e-03   1.119 0.263284
## KM:Automatic1              -1.826e-03  5.622e-03  -0.325 0.745313
## KM:CC                      -9.529e-06  2.492e-05  -0.382 0.702227
## KM:Doors                   -1.446e-03  1.334e-03  -1.085 0.278258
## KM:Weight                   8.987e-05  5.225e-05   1.720 0.085637 .
## FuelTypeDiesel:HP          -3.167e+02  1.816e+02  -1.744 0.081381 .
```

```
## FuelTypePetrol:HP                 NA         NA      NA       NA
## FuelTypeDiesel:MetColor1    3.185e+03  1.267e+03   2.515 0.012015 *
## FuelTypePetrol:MetColor1    5.459e+02  9.797e+02   0.557 0.577465
## FuelTypeDiesel:Automatic1          NA         NA      NA       NA
## FuelTypePetrol:Automatic1   3.202e+03  1.523e+03   2.102 0.035691 *
## FuelTypeDiesel:CC          -5.697e+02  8.772e+02  -0.649 0.516188
## FuelTypePetrol:CC          -6.325e+02  8.768e+02  -0.721 0.470845
## FuelTypeDiesel:Doors       -1.128e+02  7.565e+02  -0.149 0.881487
## FuelTypePetrol:Doors       -4.518e+02  5.386e+02  -0.839 0.401661
## FuelTypeDiesel:Weight       8.985e+01  2.694e+01   3.336 0.000874 ***
## FuelTypePetrol:Weight       5.585e+01  2.355e+01   2.371 0.017854 *
## HP:MetColor1                3.827e+01  1.173e+01   3.262 0.001134 **
## HP:Automatic1              -1.053e+02  1.375e+02  -0.766 0.443827
## HP:CC                       4.560e-01  3.190e-01   1.430 0.153037
## HP:Doors                   -2.100e+00  6.399e+00  -0.328 0.742807
## HP:Weight                   7.715e-01  2.066e-01   3.734 0.000196 ***
## MetColor1:Automatic1        8.829e+01  2.862e+02   0.308 0.757765
## MetColor1:CC               -2.040e+00  1.023e+00  -1.994 0.046382 *
## MetColor1:Doors             3.681e+01  7.718e+01   0.477 0.633511
## MetColor1:Weight           -6.690e+00  2.780e+00  -2.407 0.016233 *
## Automatic1:CC               7.256e+00  1.058e+01   0.686 0.492848
## Automatic1:Doors           -1.322e+02  1.683e+02  -0.785 0.432329
## Automatic1:Weight           3.104e-01  6.376e+00   0.049 0.961187
## CC:Doors                   -7.751e-01  5.741e-01  -1.350 0.177178
## CC:Weight                  -5.667e-02  2.111e-02  -2.684 0.007367 **
## Doors:Weight                3.499e+00  1.009e+00   3.467 0.000542 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1072 on 1378 degrees of freedom
## Multiple R-squared:  0.9161, Adjusted R-squared:  0.9126
## F-statistic: 263.8 on 57 and 1378 DF,  p-value: < 2.2e-16
```

In model6 we get the highest adjusted R-squared till now. Then we start to work on eliminating the insignificant variables. So, we decide to pick up those variables with smallest p-value (which is with 2 and 3 star signs) and create another model to see how adjusted R-square goes.

Model 7:

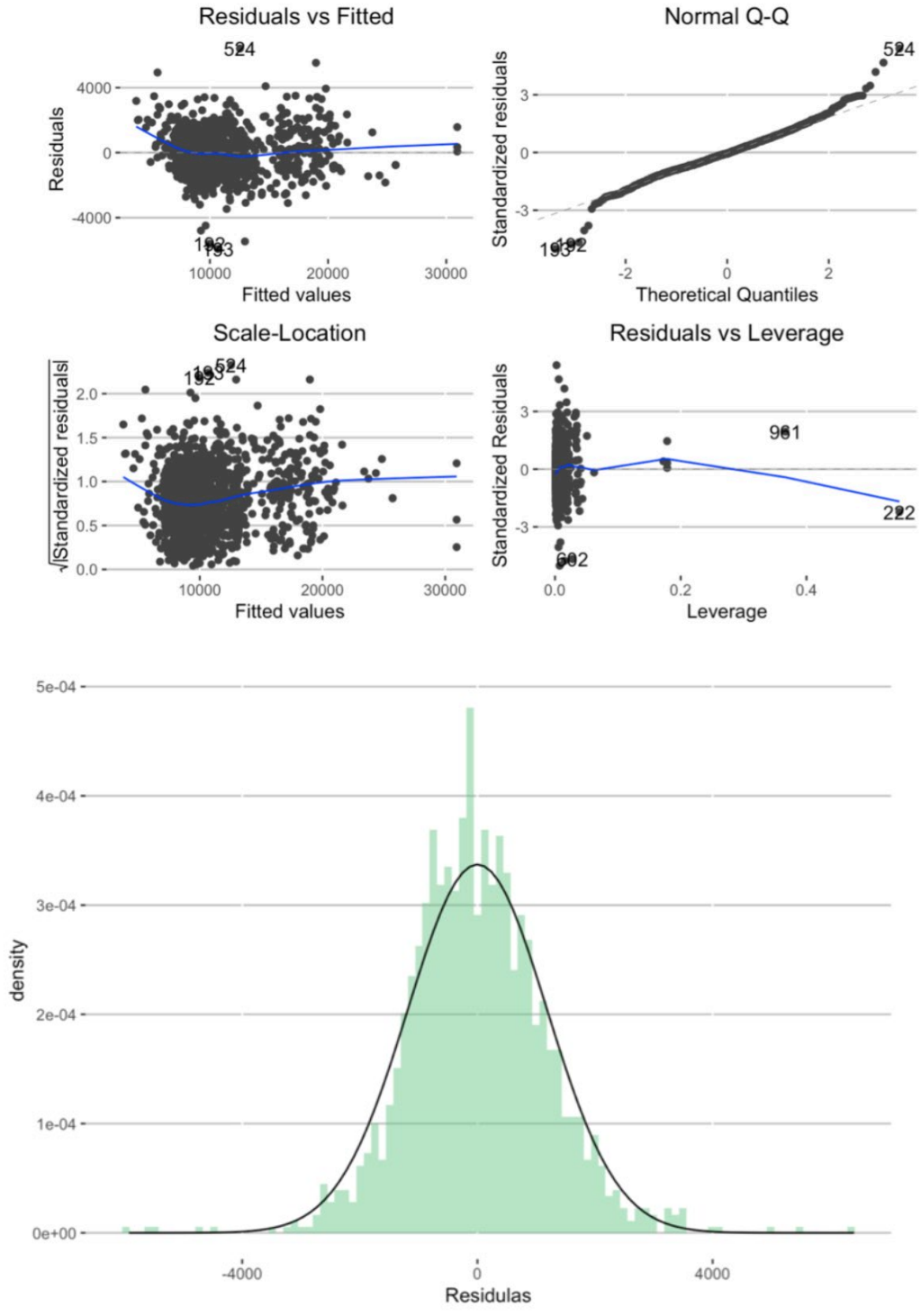
```
##
## Call:
## lm(formula = Price ~ Age + Weight + HP + I(Weight^2) + I(KM^2) +
##     I(Doors^2) + Age:CC + CC:Weight + Age:KM + Age:Doors + Age:Weight +
##     Doors:Weight, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5877.4  -755.7   -42.5   698.3  6312.8
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.057e+05  7.298e+03 -14.482  < 2e-16 ***
## Age           5.203e+02  4.041e+01  12.876  < 2e-16 ***
## Weight        1.716e+02  1.184e+01  14.487  < 2e-16 ***
## HP            2.744e+01  2.338e+00  11.736  < 2e-16 ***
## I(Weight^2)  -5.195e-02  4.901e-03 -10.600  < 2e-16 ***
## I(KM^2)      -1.182e-07  1.568e-08  -7.542 8.21e-14 ***
## I(Doors^2)   -2.314e+02  8.982e+01  -2.576 0.010084 *
## Age:CC        4.046e-02  1.256e-02   3.221 0.001307 **
## Weight:CC    -3.827e-03  6.565e-04  -5.830 6.86e-09 ***
## Age:KM        1.606e-04  4.835e-05   3.322 0.000916 ***
## Age:Doors     1.541e+00  1.903e+00   0.810 0.418344
## Age:Weight   -6.687e-01  4.852e-02 -13.781  < 2e-16 ***
## Weight:Doors  1.581e+00  6.411e-01   2.467 0.013755 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1185 on 1423 degrees of freedom
## Multiple R-squared:  0.8942, Adjusted R-squared:  0.8933
## F-statistic:  1002 on 12 and 1423 DF,  p-value: < 2.2e-16
```

The adjusted R-squared for model 7 decrease a little bit. But almost all the variables are significant. We think get rid of those useless variables will improve our model more. Especially those variables with p-value more than 0.1 and 0.01, they are too big and are insignificant. Finally, we create a new model to improve the significance and multicollinearity.

Model 8:

```
##
## Call:
## lm(formula = Price ~ Age + Weight + HP + I(Weight^2) + I(KM^2) +
##     Age:CC + CC:Weight + Age:KM + Age:Weight, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5926.7  -765.3   -64.2   712.1  6405.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.014e+05  7.004e+03 -14.478  < 2e-16 ***
## Age          5.604e+02  3.736e+01  14.999  < 2e-16 ***
## Weight       1.633e+02  1.125e+01  14.513  < 2e-16 ***
## HP           2.718e+01  2.328e+00  11.676  < 2e-16 ***
## I(Weight^2) -4.546e-02  4.281e-03 -10.619  < 2e-16 ***
## I(KM^2)     -1.145e-07  1.564e-08  -7.320 4.13e-13 ***
## Age:CC       4.728e-02  1.204e-02   3.926 9.06e-05 ***
## Weight:CC   -3.849e-03  6.451e-04  -5.967 3.05e-09 ***
## Age:KM       1.510e-04  4.830e-05   3.127   0.0018 **
## Age:Weight  -7.104e-01  4.340e-02 -16.369  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1187 on 1426 degrees of freedom
## Multiple R-squared:  0.8936, Adjusted R-squared:  0.8929
## F-statistic:  1331 on 9 and 1426 DF,  p-value: < 2.2e-16
```

The model8 is our final model. Based on the model7 we did before, we delete the insignificant variables and keep the significant variables. Now almost all variables are highly correlated (low p-value) and the final adjusted R-squared we get is 0.8929, which is quite high. From our final model, we can conclude that for this model, price is not only interpreted by the single variables age, Weight and HP, but also some other variables like CC, KM are used in addition to some interactions and square terms. The interpretation of these interactions is hard to explain, and might cannot be used to value in real life. However, it's valuable to keep it.



Diagnostic plots above were created to verify the linear regression model assumptions. Although there are still some outliers in the Residual plot and also in Q-Q plot, the overall plot shown up a linear trend and most points fit well. There are only several points for the outliers, so it won't influence the whole model that much. The bell-shaped pattern of a histogram further proved that the model is a well normal distributed one.

Question1:

From the data and model we got, the variables significant to sales price are:

- Age
- Weight
- HP
- Square of Weight
- Square of KM
- Interaction between Age and CC
- Interaction between Age and KM
- Interaction between Age and Weight
- Interaction between Weight and CC

Question2:

Although in the model we got before, we already knew that Age is an important factor that will influence the price of car. Our manager believes that a customer would pay more, and in particular, 1000 Euros more if customers are 10 years older. However, it's still hard to say there's a 100 times Euros increase for 1 year increase in age of purchaser. So we apply the simple linear regression to predict the sales prices only with the ages.

```
Call:
lm(formula = Price ~ Age, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8423.0  -997.4   -24.6   878.5 12889.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 20294.059    146.097  138.91   <2e-16 ***
Age          -170.934      2.478  -68.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1746 on 1434 degrees of freedom
Multiple R-squared:  0.7684,    Adjusted R-squared:  0.7682
F-statistic:  4758 on 1 and 1434 DF,  p-value: < 2.2e-16

                 2.5 %     97.5 %
(Intercept) 20007.4714 20580.6459
Age          -175.7946  -166.0725
```

To formulate this problem as a hypothesis test, the null hypothesis is that for 1 year increase in Age of purchaser will lead to 100 Euros increase in Sales Price. (1000Euros more for 10 years older, so 100 Euros more for 1 year older). Therefore, the equivalent hypothesis is that the Age coefficient will be 100 Euros. Alternatively, the coefficient is not 100 Euros.

$$H_0: \beta_1 = 100$$

$$H_0 : \beta_1 = 100$$

$$H_1: \beta_1 \neq 100$$

$$H_1 : \beta_1 \neq 100$$

A 95% confidence interval for the Age coefficient is such that the coefficient is between -175.795Euros and -166.073 Euros. Since 100 Euros is outside the confidence interval, we should reject the null hypothesis based on a 95% level of significance. We can also observe that the coefficient of Age (which is -170.934 Euros) is significantly different than 100 Euros.

Question 3:

The manager also believed that customers typically pay an additional 3000 Euros for an automatic transmission. As a dummy variable, Automatic shows up this car is automatic or manual. To examine whether this hypothesis is true, we also linear regression model to fit the Price with the dummy variable Automatic and find the coefficient.

```
Call:
lm(formula = Price ~ Automatic, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6351.7 -2274.6  -804.2  1248.3 21798.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10701.69      98.48 108.674   <2e-16 ***
Automatic1    522.93     417.21   1.253     0.21
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3626 on 1434 degrees of freedom
Multiple R-squared:  0.001094,  Adjusted R-squared:  0.0003977
F-statistic: 1.571 on 1 and 1434 DF,  p-value: 0.2103

                 2.5 %    97.5 %
(Intercept) 10508.5208 10894.863
Automatic1   -295.4827  1341.349
```

To formulate this problem as a hypothesis test, the null hypothesis is that for Automatic transmission is 3000Euros (as Automatic=1). Therefore, the equivalent hypothesis is that the Automatic coefficient will be 3000Euros. Alternatively, the coefficient is not 3000Euros.

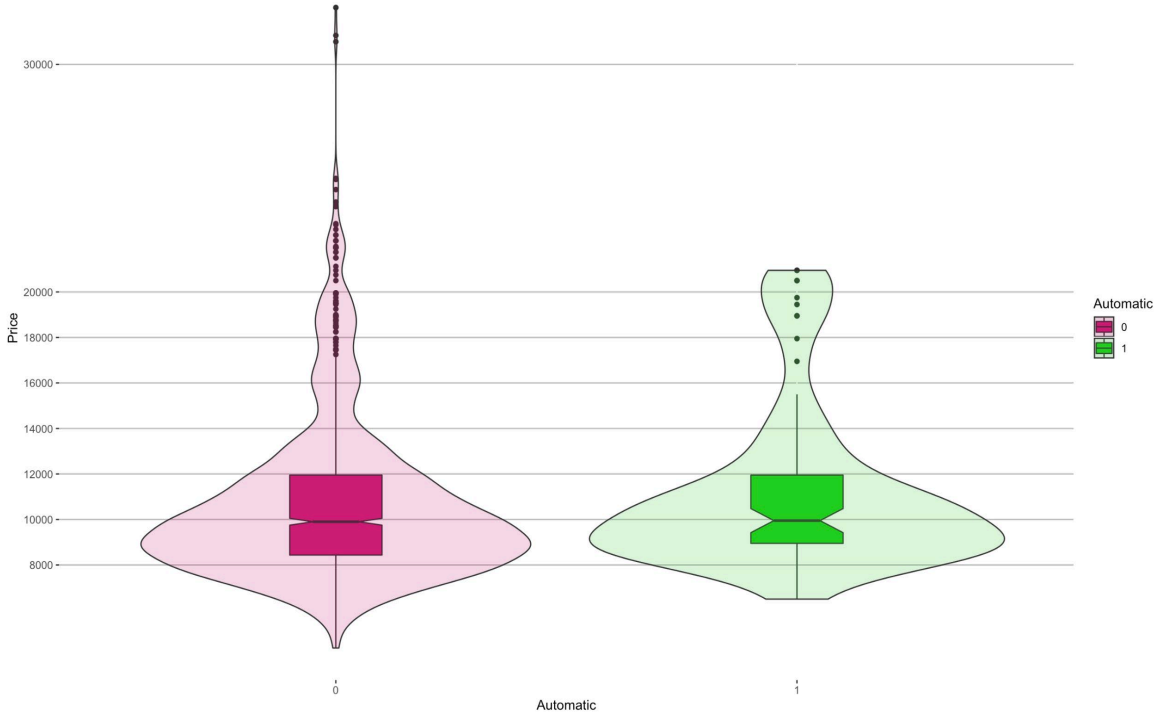
$$H_0: \beta_1 = 3000$$

$$H_0 : \beta_1 = 3000$$

$$H_1: \beta_1 \neq 3000$$

$$H_1 : \beta_1 \neq 3000$$

A 95% confidence interval for the Automatic coefficient is such that the coefficient is between -295.483Euros and 1341.349Euros. Since 3000Euros is outside the confidence interval, we should reject the null hypothesis based on a 95% level of significance. We can also observe that the coefficient of Automatic (which is 522.92Euros) is significantly less than 3000Euros.



We also drawn a boxplot of sales prices for dummy variable Automatic. From observing the Figure above, it shows that the value in Manual (Automatic=0) is larger and had a wider spread for the price. The midpoint of Automatic and Manual is similar(around 10000Euros) So there's no evidence that price of Automatic transmission car is 3000 Euros higher than Manual Transmission car.

Conclusion

In the analyzation we did, a regression model that contains interaction between two variables and square of variables was used to interpret the relationships between prices and all other variables. We started from the simple linear regression first and continuously improve the model by try the method of adding interaction between 2

variables, the square of numeric variables. Finally, the linear regression model could fit the dataset and interpreted the significance of all variable used as indicated in the adjusted R-square of 0.8929.

Another question about relationship between Age and Price was added. Our manager estimated that there was a 100 times relationship between Price and Age, as Manager claimed that 1000 Euros more on Price can be ascribed to the 10 years older for the purchaser. However, through our analysis and model, we found that value of 100 Euros was far outside the 95% confidence interval of the coefficient for Age. Even the variable Age itself is negatively correlated with Price.

The additional prediction was that if the car has Automatic transmission, there will be a 3000 Euros more on price. However, through our linear regression models, we proved that the 3000 Euros is outside the 95% confidence interval of coefficient for Automatic. So the hypothesis is not true.

Appendix

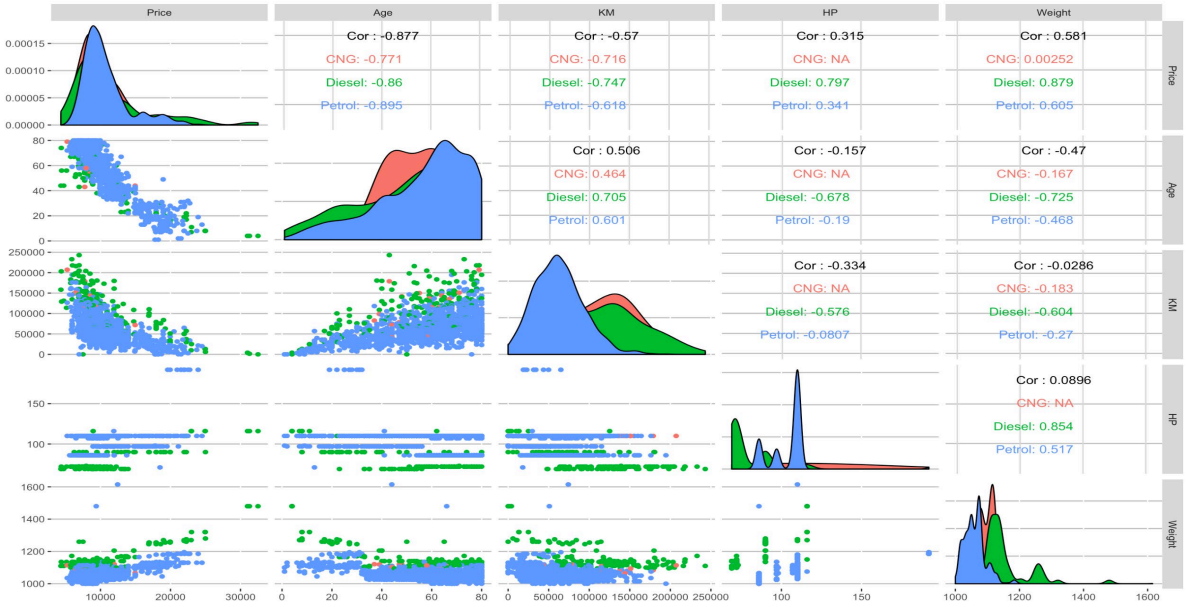


Figure2

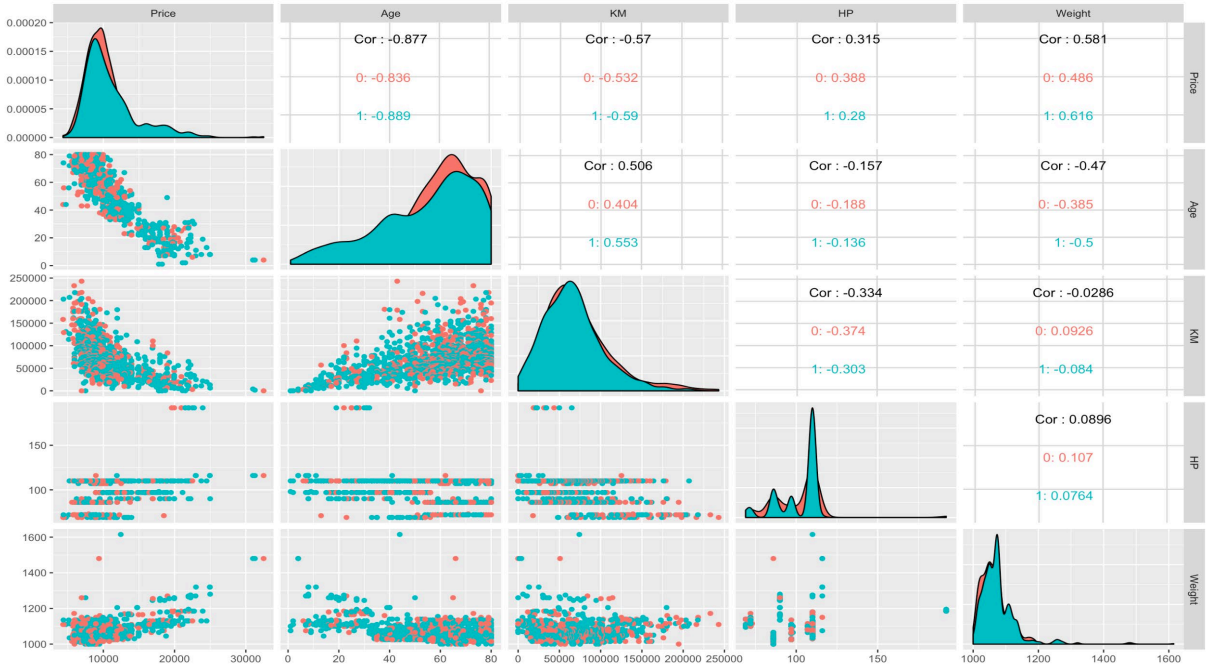


Figure 3

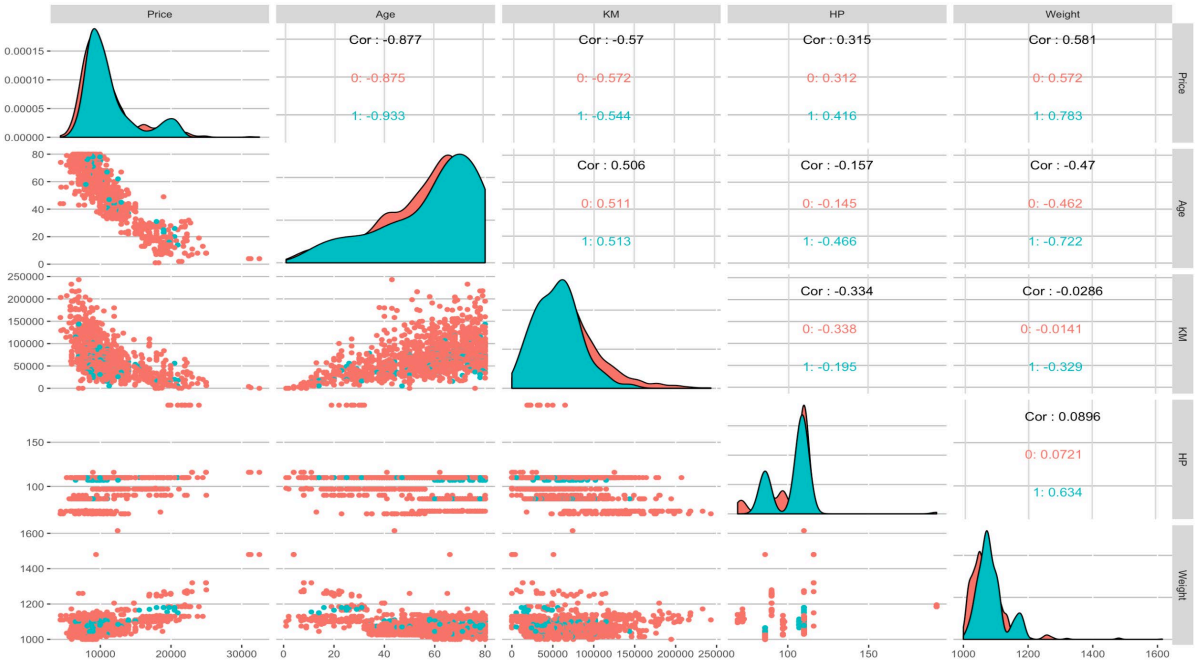

Price
Age
KM
HP
Weight
Cor : -0.877
0: -0.875
1: -0.933
Cor : -0.57
0: -0.572
1: -0.544
Cor : 0.315
0: 0.312
1: 0.416
Cor : 0.581
0: 0.572
1: 0.783
Cor : 0.506
0: 0.511
1: 0.513
Cor : -0.157
0: -0.145
1: -0.466
Cor : -0.47
0: -0.462
1: -0.722
Cor : -0.334
0: -0.338
1: -0.195
Cor : -0.0286
0: -0.0141
1: -0.329
Cor : 0.0896
0: 0.0721
1: 0.634

Figure4