Used car website applications analysis

Introduction

Sales forecasting and market trends prediction are one of the most popular topics nowadays for sellers and buyers. The modeling may help to reduce time for price definition and rely on algorithms. Moreover, it is possible to understand which characteristics influence the most. At this report the used car sales are explored to see the trends, demand and what features influence on the price the most.

Data analysis

Dataset overview

Car market is in demand now. There are also websites for used vehicles. Thus, the website drom.ru proposes the services for people, who would like to sell their cars. The data contains the information about the applications from Ekaterinburg city (Russia). The closer overview of the dataset follows. There is a structure of the data:

```
## 'data.frame':    11136 obs. of  15 variables:
##  $ firm             : chr  "Ford" "Hyundai" "Toyota" "Suzuki" ...
##  $ model            : chr  "Fiesta" "Santa Fe Classic" "RAV4" "Grand Vitara XL-7" ...
##  $ drive            : chr  "FF" "FF" "4WD" "4WD" ...
##  $ frame_type       : chr  "HATCH" "SUV" "SUV" "SUV" ...
##  $ generation_number: int  5 1 3 -1 1 1 1 6 3 1 ...
##  $ year             : int  2007 2008 2010 2007 2008 2008 2002 2007 2007 2012 ...
##  $ restyling_number : int  1 0 1 -1 0 0 0 0 0 2 ...
##  $ mileage          : int  93000 78000 97000 77000 75000 97000 130000 96000 138000 50000 ...
##  $ transmission     : chr  "MANUAL" "MANUAL" "AUTO" "AUTO" ...
##  $ engine_power     : int  101 112 152 250 140 238 234 95 190 80 ...
##  $ steering_wheel   : chr  "left" "left" "left" "left" ...
##  $ fuel_type        : chr  "GASOLINE" "DIESEL" "GASOLINE" "GASOLINE" ...
##  $ volume           : num  1600 2000 2000 3600 2400 2300 3500 1300 2700 1500 ...
##  $ ownership_periods: num  4 1 2 5 2 2 7 3 3 1 ...
##  $ price            : int  258000 495000 1050000 755000 640000 565000 445000 295000 815000 248000 ..
```

The data contains cars' features, price, owners' comments and specific information for the website. Some values are missing, but only few, therefore it is possible to continue the analysis. One of the rows, which has all the information:

```
##   firm  model drive frame_type generation_number year restyling_number mileage
## 1 Ford Fiesta    FF      HATCH                 5 2007                1   93000
##   transmission engine_power steering_wheel fuel_type volume ownership_periods
## 1       MANUAL          101           left  GASOLINE   1600                 4
##    price
## 1 258000
```

Prelimenary analysis

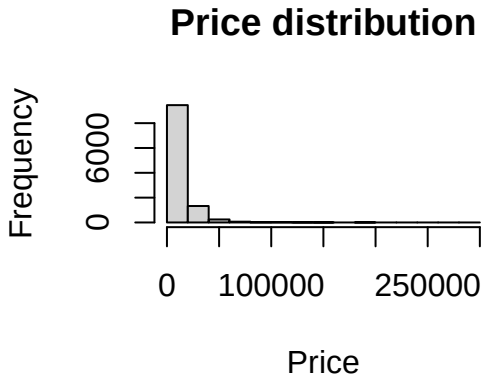
The dataset has been checked for unknowns and wrong data. They have been changed on NAs. The price is in rubles, a new column in euros has been added. Thus, the minimum, maximum and average values of price in euros were found.

```
##    average    max min
## 1 11994.71 295443 341
```

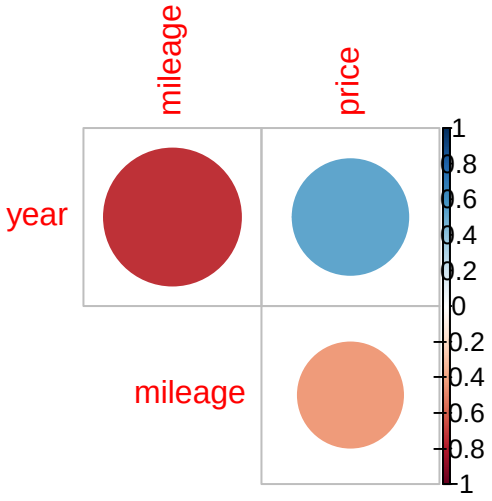
The side of steering wheel was considered separately. The T- and F-tests were applied to see, how these two categories differ from each other. P-values show, that mean and variance differ significantly, because they are less than 0.05. It is explained by high demand on left-side steering wheels and low diversity of right side steering wheel cars.

```
##
##  F test to compare two variances
##
## data:  left$price_euro and right$price_euro
## F = 7.5186, num df = 10357, denom df = 745, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  6.751361 8.332449
## sample estimates:
## ratio of variances
##           7.518586
```

```
##
##  Welch Two Sample t-test
##
## data:  left$price_euro and right$price_euro
## t = 32.624, df = 1733.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  6730.361 7591.384
## sample estimates:
## mean of x mean of y
##  12476.32   5315.45
```



The left histogram shows, that the most cars (10145) cost 25 thousand euro or less, however, there are 1008, which shift the mean. After the log function transformation the histogram (right) has changed to a normal distribution. Therefore it may be Poisson distribution.



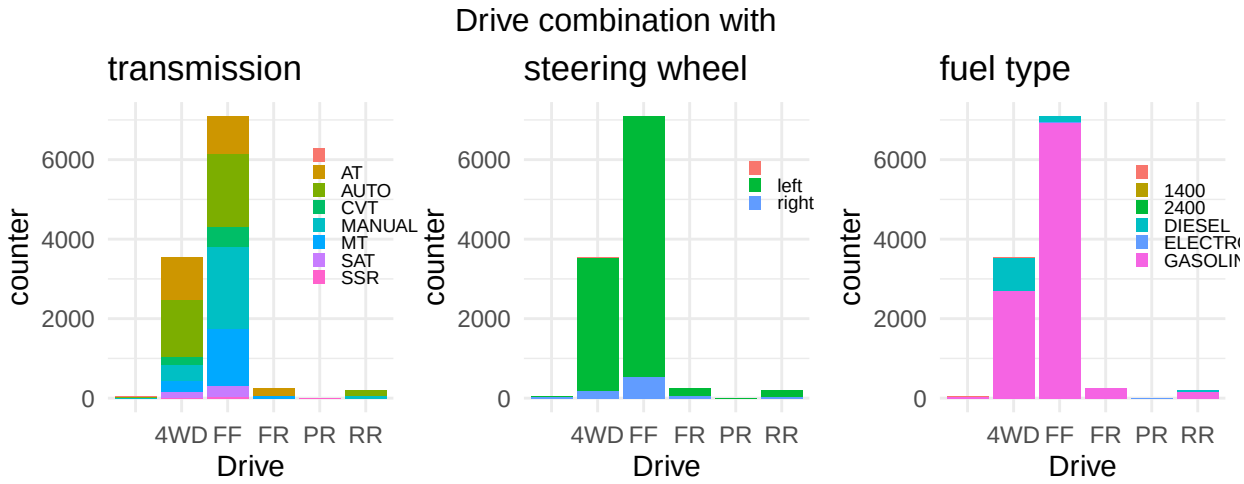
The correlation plot helps to understand the dependency between variables. It can be seen that 'year' and 'mileage' have opposite correlation coefficient. Indeed, earlier year of the car release matches more mileage. On the other hand, the 'price' cannot be predicted fully by 'year' and 'mileage'. Also it can be explained by appearance of retro cars, when the price is high for the old vehicle. The earliest year of car is 1972.

Applications' trends

The data is diverse, many features are skewed to some preferences, the price depends not always directly on characteristics, but also on the status of the cars.

The most popular model is Focus of Ford: 290 items.

The four common features for buyers are drive, transmission, steering wheel and fuel types. They are categorical values and may show customer preferences.



The transmission types on drives are diverse, therefore it is hard to say what is the most popular. Nonetheless, some of them are more frequent. The domination of the left side steering wheel is shown on the middle graph. The gasoline is more usual fuel type.

The dataset analysis revealed the observation of applications, preferences. But what eventually defines the price and why even used cars cost expensive? Based on data the model will be design to understand how the price is predicted. Let me consider it closely at the next section.

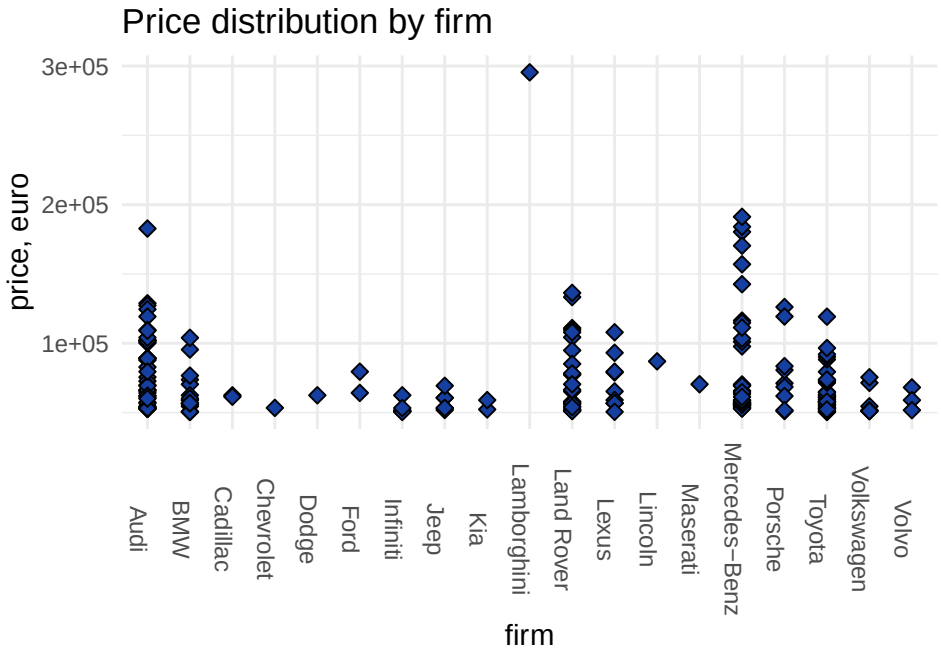
Price definition

New data creation

After data analysis and plots there is still a question: How price is defined? It was decided to investigate the most expensive cars and realize, how the cost is built. The most expensive cars, namely the price >50 000 euro, were selected for further analysis.

New table has 202 rows and diverse from initial one: now the steering wheel is only at the left, the drive is 4WD, except one FF.

In order to understand the price distribution the plot by variable 'firm' was created.



We can see one outlier: it is Lamborghini, even all other cars cost below 20 thousand euros.

Model design

Categorical explanatory variables, not normally distributed data lead us to use mixed effect models to predict the car price. The main idea is to find the best model, which can forecast the final price based on definite characteristics. Models are compared by AIC.

At first, only firms and models of cars were taken into consideration.

The firm and model influence a lot on the decision of the price. But it is not always the case, because the cost also depends on the conditions and features. The mixed models with 'firm' and 'model' have quite high AICs: 4717.66, 4655.07. Now we can restrict the model for other variables.

The initial model contains all the variables without interactions. We can see by value, which predictors we can save for further work.

The summary of this and follows models are presented at the Appendix.

Initial model

AIC of initial model is 4614.27. Mileage, engine power, generation number, year and ownership periods need to be included at this step. Also type of fuel influences on the price. Now we simplify the model.

1st model

AIC of the 1st model is 4608.88. Ownership periods at this step don't have enough weight anymore. We can exclude it.

2nd model

AIC of the 2nd model is 4609.54. Now we have only 5 variables, therefore we may create interactions, excluding more than 2.

3rd model

AIC of the 3rd model is 4550.65. We can exclude the fuel type from the model.

4th model

AIC of the 3rd model is 4585.84. We can see that AIC became higher, therefore, it is better to modify the previous model. Thus, all 5 variables 'mileage', 'engine power', 'generation number', 'year' and 'fuel type' will be at the model and several interactions, which had significant impact will be added to the 5th model.

5th model

AIC of the 3rd model is 4548.08. AIC of this model is the lowest and all terms are significant (mileage and engine power, mileage and year, engine power and year). We can stop here and choose this model, that explains the car price in best way.

Conclusion

At this report the data from website of used car applications has been explored. The analysis showed the most frequent combinations of features, popular models, price distribution. For the price policy of most expensive cars understanding the model has been found. The initial model contained all the variables and later has been simplified. The most appropriate model has been chosen by AIC and less number of variables. Thus, the 'mileage', 'engine power', 'generation number', 'year' and 'fuel type' are more significant in prediction the price of most expensive cars.

Appendix

Models‘ summaries

Initial model

```
##
## Call:
## lm(formula = focus$price_euro ~ focus$frame_type + focus$fuel_type +
##     focus$transmission + focus$mileage + focus$engine_power +
##     focus$generation_number + focus$volume + focus$year + focus$ownership_periods +
##     focus$restyling_number)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -51047 -10908     38   8324 126187
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.062e+07  2.744e+06  -3.868 0.000153 ***
## focus$frame_typeHATCH      -4.132e+03  1.661e+04  -0.249 0.803799
## focus$frame_typeLIFTBACK    1.185e+03  1.261e+04   0.094 0.925239
## focus$frame_typeMINIVAN     7.455e+03  1.763e+04   0.423 0.672941
## focus$frame_typePICKUP     -2.660e+04  1.556e+04  -1.710 0.088996 .
## focus$frame_typeSEDAN       1.728e+04  1.061e+04   1.629 0.105144
## focus$frame_typeSUV         2.749e+03  9.973e+03   0.276 0.783119
## focus$frame_typeSUV_3DR     1.065e+04  1.781e+04   0.598 0.550627
## focus$fuel_typeELECTRO     -3.500e+04  1.371e+04  -2.553 0.011516 *
## focus$fuel_typeGASOLINE    -2.000e+04  3.725e+03  -5.367 2.45e-07 ***
## focus$transmissionAUTO     -2.248e+03  3.177e+03  -0.708 0.480106
## focus$transmissionCVT      -1.881e+04  1.527e+04  -1.232 0.219452
## focus$transmissionMANUAL    9.019e+03  2.200e+04   0.410 0.682378
## focus$transmissionMT        3.095e+04  2.183e+04   1.418 0.158013
## focus$transmissionSAT      -2.265e+04  1.299e+04  -1.743 0.083027 .
## focus$transmissionSSR      -2.836e+04  2.384e+04  -1.190 0.235785
## focus$mileage              -2.632e-01  6.970e-02  -3.776 0.000216 ***
## focus$engine_power          2.968e+02  2.722e+01  10.906  < 2e-16 ***
## focus$generation_number     2.417e+03  5.992e+02   4.034 8.10e-05 ***
## focus$volume               -1.936e+00  2.364e+00  -0.819 0.414024
## focus$year                  5.259e+03  1.358e+03   3.872 0.000151 ***
## focus$ownership_periods    -3.907e+03  1.928e+03  -2.026 0.044237 *
## focus$restyling_number     -1.979e+02  2.128e+03  -0.093 0.926003
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 20830 on 179 degrees of freedom
## Multiple R-squared:  0.6221, Adjusted R-squared:  0.5757
## F-statistic:  13.4 on 22 and 179 DF,  p-value: < 2.2e-16
```

1st model

```
##
## Call:
## lm(formula = focus$price_euro ~ focus$fuel_type + focus$mileage +
##     focus$engine_power + focus$generation_number + focus$year +
##     focus$ownership_periods)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45465 -11624  -1333   8319 131967
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.123e+07  2.570e+06  -4.367 2.05e-05 ***
## focus$fuel_typeELECTRO   -3.051e+04  9.909e+03  -3.079  0.00238 **
## focus$fuel_typeGASOLINE -2.035e+04  3.565e+03  -5.707 4.25e-08 ***
## focus$mileage            -3.075e-01  6.734e-02  -4.566 8.82e-06 ***
## focus$engine_power        2.715e+02  1.952e+01  13.909  < 2e-16 ***
## focus$generation_number   2.289e+03  5.037e+02   4.544 9.69e-06 ***
## focus$year                5.563e+03  1.272e+03   4.374 1.99e-05 ***
## focus$ownership_periods -3.052e+03  1.905e+03  -1.602  0.11079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21270 on 194 degrees of freedom
## Multiple R-squared:  0.5732, Adjusted R-squared:  0.5578
## F-statistic: 37.22 on 7 and 194 DF,  p-value: < 2.2e-16
```

2nd model

```
##
## Call:
## lm(formula = focus$price_euro ~ focus$fuel_type + focus$mileage +
##     focus$engine_power + focus$generation_number + focus$year)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47610 -11977  -1423   8529 129247
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.302e+07  2.324e+06  -5.602 7.14e-08 ***
## focus$fuel_typeELECTRO   -2.942e+04  9.926e+03  -2.964  0.00342 **
## focus$fuel_typeGASOLINE -1.978e+04  3.562e+03  -5.553 9.12e-08 ***
## focus$mileage            -3.187e-01  6.724e-02  -4.740 4.12e-06 ***
## focus$engine_power        2.665e+02  1.935e+01  13.775  < 2e-16 ***
## focus$generation_number   2.293e+03  5.057e+02   4.534 1.01e-05 ***
## focus$year                6.450e+03  1.150e+03   5.610 6.87e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21350 on 195 degrees of freedom
## Multiple R-squared:  0.5675, Adjusted R-squared:  0.5542
## F-statistic: 42.65 on 6 and 195 DF,  p-value: < 2.2e-16
```

3rd model

```
##
## Call:
## lm(formula = focus$price_euro ~ (focus$fuel_type + focus$mileage +
##     focus$engine_power + focus$generation_number + focus$year)^2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55077  -9055  -1365   5370  91752
##
## Coefficients: (1 not defined because of singularities)
##                                                Estimate Std. Error t value
## (Intercept)                                   1.280e+07  8.628e+06   1.483
## focus$fuel_typeELECTRO                        4.738e+04  7.267e+04   0.652
## focus$fuel_typeGASOLINE                       9.561e+06  5.170e+06   1.849
## focus$mileage                                 3.292e+02  5.935e+01   5.546
## focus$engine_power                           -1.211e+05  2.815e+04  -4.302
## focus$generation_number                      -6.838e+05  6.541e+05  -1.045
## focus$year                                   -6.342e+03  4.270e+03  -1.485
## focus$fuel_typeELECTRO:focus$mileage         -5.940e-01  2.644e+00  -0.225
## focus$fuel_typeGASOLINE:focus$mileage         3.368e-02  1.510e-01   0.223
## focus$fuel_typeELECTRO:focus$engine_power    -2.333e+02  1.778e+02  -1.312
## focus$fuel_typeGASOLINE:focus$engine_power    1.434e+01  5.081e+01   0.282
## focus$fuel_typeELECTRO:focus$generation_number -5.685e+02  2.075e+04  -0.027
## focus$fuel_typeGASOLINE:focus$generation_number  2.366e+01  1.112e+03   0.021
## focus$fuel_typeELECTRO:focus$year                    NA         NA      NA
## focus$fuel_typeGASOLINE:focus$year           -4.744e+03  2.558e+03  -1.855
## focus$mileage:focus$engine_power             -1.825e-03  8.071e-04  -2.261
## focus$mileage:focus$generation_number        -1.625e-02  1.656e-02  -0.981
## focus$mileage:focus$year                     -1.629e-01  2.939e-02  -5.542
## focus$engine_power:focus$generation_number    3.640e-01  8.313e+00   0.044
## focus$engine_power:focus$year                 6.012e+01  1.393e+01   4.315
## focus$generation_number:focus$year            3.395e+02  3.237e+02   1.049
##                                              Pr(>|t|)
## (Intercept)                                    0.1397
## focus$fuel_typeELECTRO                         0.5152
## focus$fuel_typeGASOLINE                        0.0661 .
## focus$mileage                                1.02e-07 ***
## focus$engine_power                           2.76e-05 ***
## focus$generation_number                        0.2973
## focus$year                                     0.1392
## focus$fuel_typeELECTRO:focus$mileage           0.8225
## focus$fuel_typeGASOLINE:focus$mileage          0.8237
## focus$fuel_typeELECTRO:focus$engine_power      0.1913
## focus$fuel_typeGASOLINE:focus$engine_power     0.7781
## focus$fuel_typeELECTRO:focus$generation_number  0.9782
## focus$fuel_typeGASOLINE:focus$generation_number  0.9830
## focus$fuel_typeELECTRO:focus$year                 NA
## focus$fuel_typeGASOLINE:focus$year             0.0652 .
## focus$mileage:focus$engine_power               0.0249 *
## focus$mileage:focus$generation_number          0.3279
## focus$mileage:focus$year                     1.03e-07 ***
## focus$engine_power:focus$generation_number     0.9651
```

```
## focus$engine_power:focus$year                    2.62e-05 ***
## focus$generation_number:focus$year                 0.2957
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17910 on 182 degrees of freedom
## Multiple R-squared:  0.7159, Adjusted R-squared:  0.6863
## F-statistic: 24.14 on 19 and 182 DF,  p-value: < 2.2e-16
```

4th model

```
##
## Call:
## lm(formula = focus$price_euro ~ (focus$mileage + focus$engine_power +
##     focus$generation_number + focus$year)^2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -40815 -11520  -2771   6334 111401
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                               9.628e+06  9.477e+06   1.016
## focus$mileage                             2.660e+02  6.230e+01   4.270
## focus$engine_power                       -8.009e+04  2.570e+04  -3.116
## focus$generation_number                  -9.907e+05  7.137e+05  -1.388
## focus$year                               -4.772e+03  4.690e+03  -1.017
## focus$mileage:focus$engine_power         -1.804e-03  7.230e-04  -2.496
## focus$mileage:focus$generation_number    -1.618e-02  1.790e-02  -0.904
## focus$mileage:focus$year                 -1.316e-01  3.086e-02  -4.264
## focus$engine_power:focus$generation_number -1.215e+00  6.814e+00  -0.178
## focus$engine_power:focus$year             3.980e+01  1.272e+01   3.129
## focus$generation_number:focus$year        4.919e+02  3.532e+02   1.393
##                                          Pr(>|t|)
## (Intercept)                               0.31092
## focus$mileage                            3.08e-05 ***
## focus$engine_power                        0.00211 **
## focus$generation_number                   0.16671
## focus$year                                0.31021
## focus$mileage:focus$engine_power          0.01341 *
## focus$mileage:focus$generation_number     0.36726
## focus$mileage:focus$year                 3.15e-05 ***
## focus$engine_power:focus$generation_number  0.85864
## focus$engine_power:focus$year             0.00203 **
## focus$generation_number:focus$year        0.16537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19950 on 191 degrees of freedom
## Multiple R-squared:  0.6303, Adjusted R-squared:  0.611
## F-statistic: 32.57 on 10 and 191 DF,  p-value: < 2.2e-16
```

5th model

```
##
## Call:
## lm(formula = focus$price_euro ~ focus$fuel_type + focus$mileage +
##     focus$engine_power + focus$generation_number + focus$year +
##     focus$mileage:focus$engine_power + focus$mileage:focus$year +
##     focus$engine_power:focus$year)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47758 -10840  -1824   6399  94147
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       8.043e+06  7.227e+06   1.113 0.267161
## focus$fuel_typeELECTRO           -4.956e+04  8.784e+03  -5.642 5.95e-08 ***
## focus$fuel_typeGASOLINE          -1.666e+04  3.063e+03  -5.441 1.60e-07 ***
## focus$mileage                     2.282e+02  5.221e+01   4.371 2.02e-05 ***
## focus$engine_power               -8.405e+04  2.242e+04  -3.749 0.000235 ***
## focus$generation_number           1.347e+03  4.448e+02   3.029 0.002794 **
## focus$year                       -3.987e+03  3.577e+03  -1.115 0.266415
## focus$mileage:focus$engine_power -1.365e-03  6.119e-04  -2.231 0.026839 *
## focus$mileage:focus$year         -1.130e-01  2.588e-02  -4.368 2.05e-05 ***
## focus$engine_power:focus$year     4.178e+01  1.110e+01   3.765 0.000221 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18210 on 192 degrees of freedom
## Multiple R-squared:  0.6903, Adjusted R-squared:  0.6758
## F-statistic: 47.55 on 9 and 192 DF,  p-value: < 2.2e-16
```