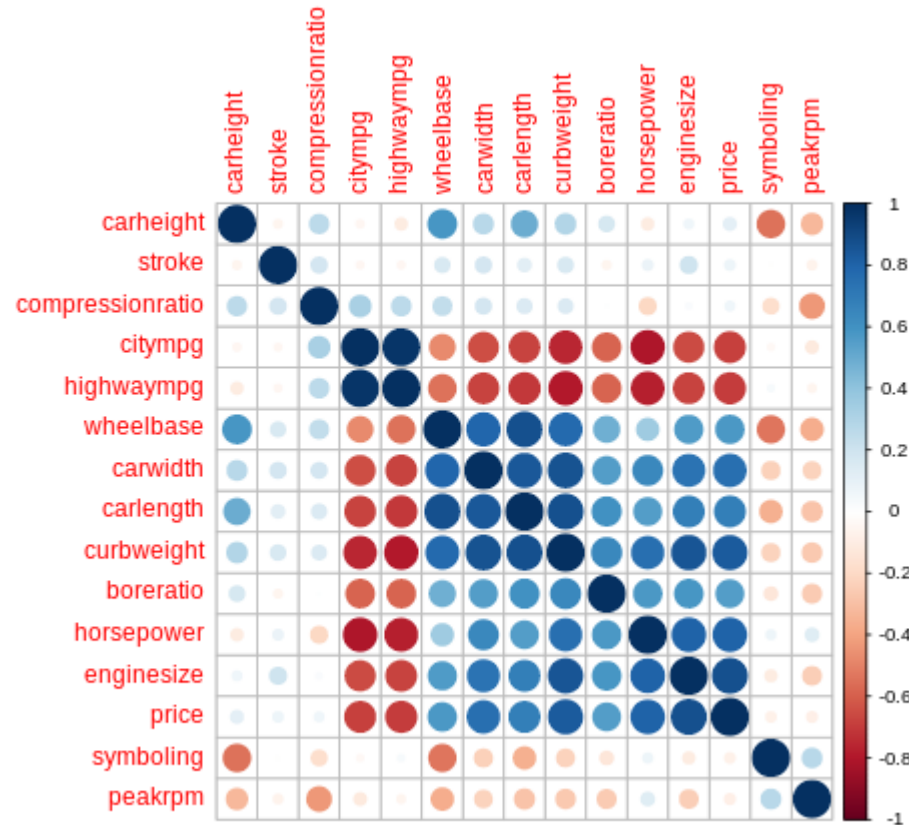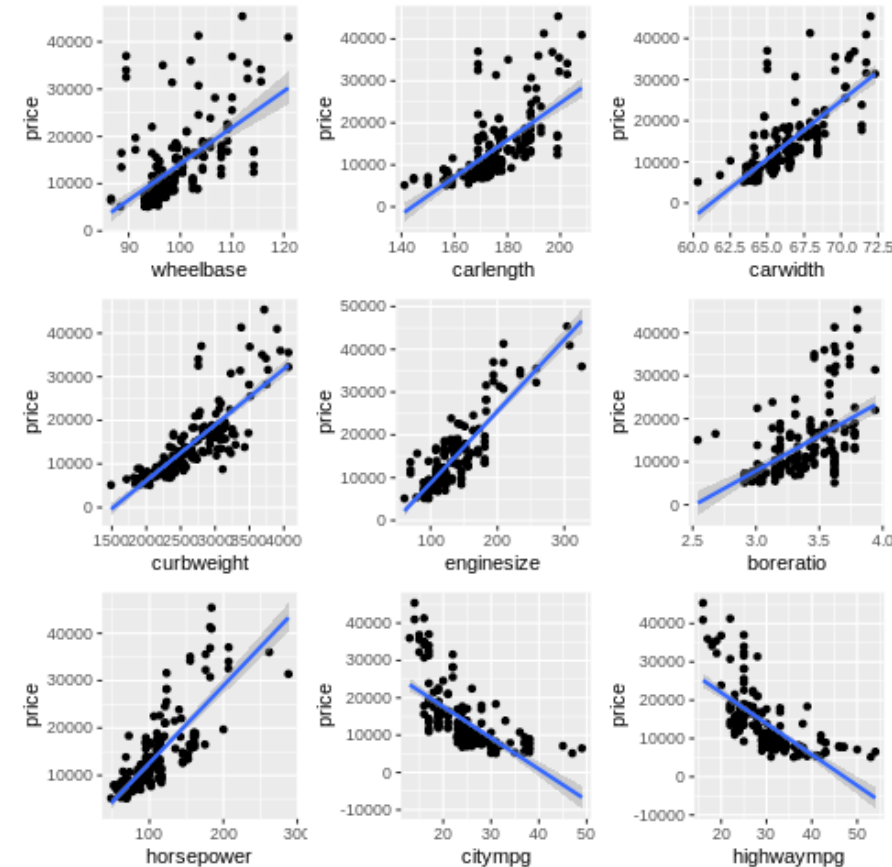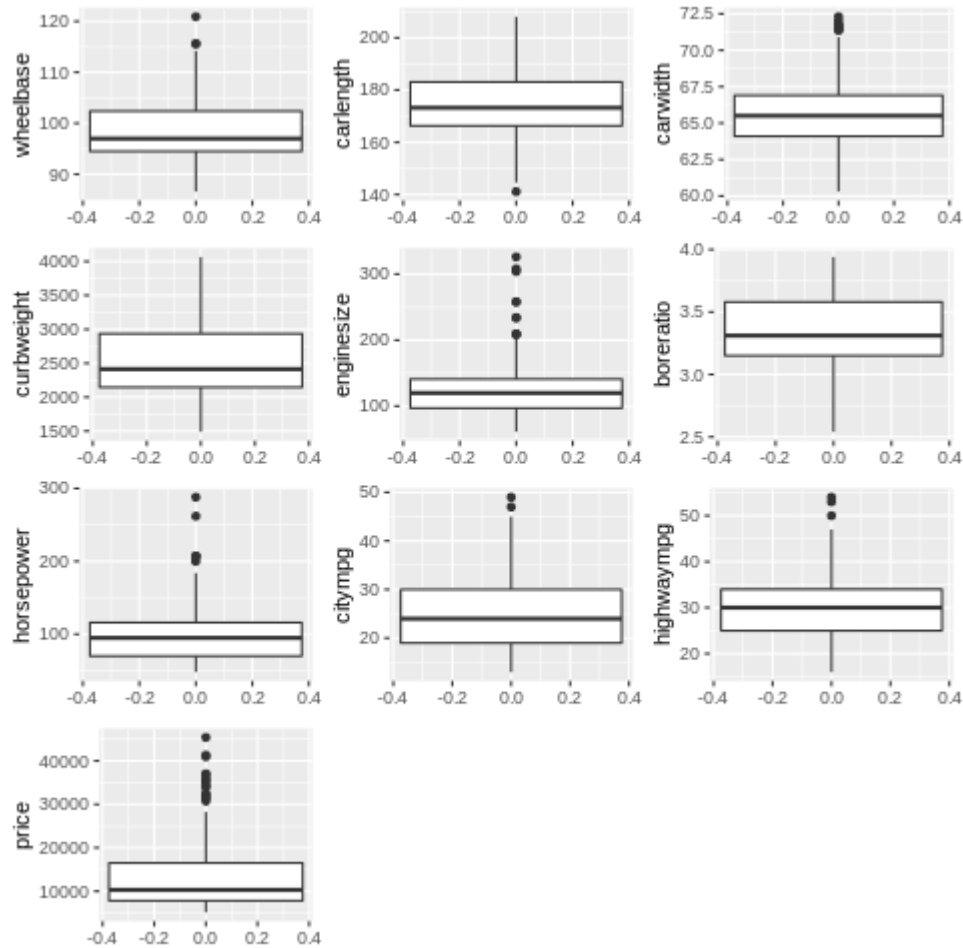# Car Price Prediction - EDA

There are several variables highly correlated with Price, with some multicollinearity amongst those variables
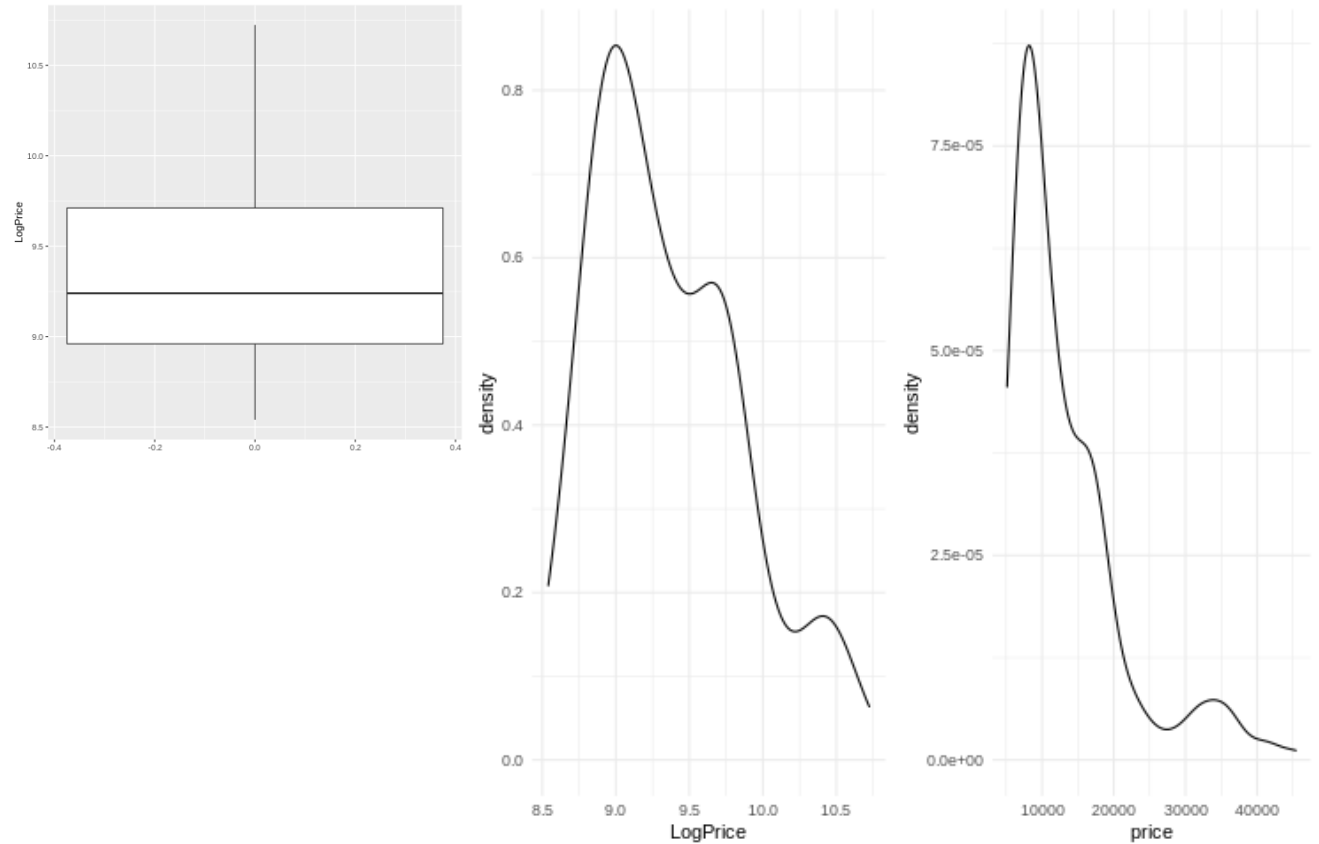
Additionally, there are some non-linear trends that could benefit from polynomial fits
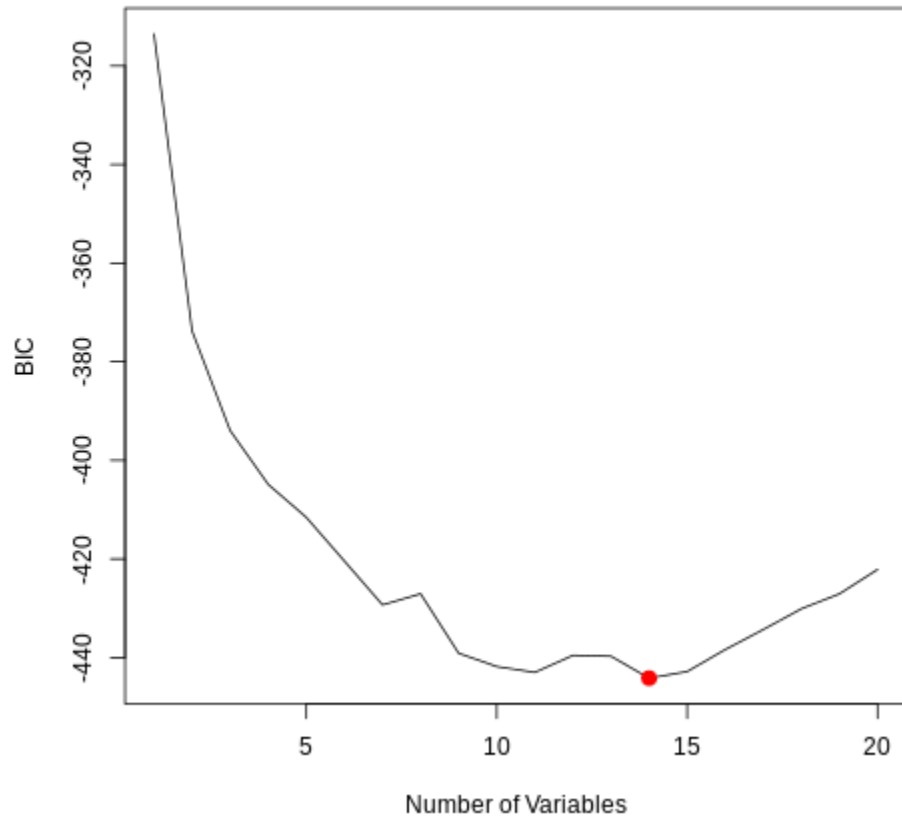
EDA – Cont.

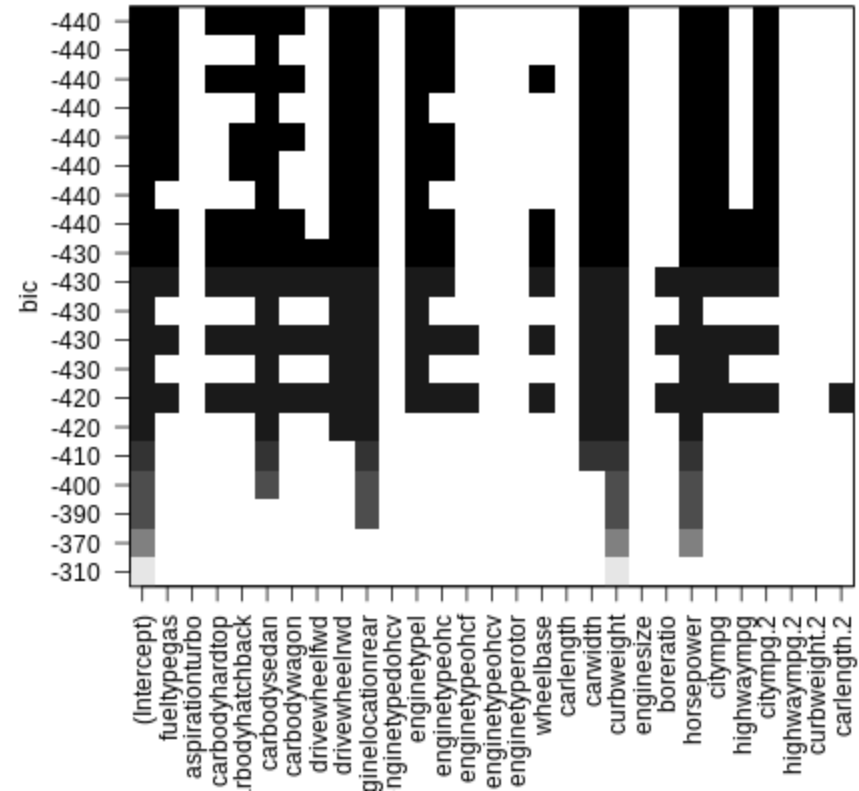Distributions were plotted to identify potential outliers

From this, Price was Log transformed to reduce skew.

Step-Wise Selection

The data was broken into training and testing data, and BIC was used with forward selection to optimize the choice of model features

The above plot show the selection process

Linear Model

```
Residuals:
     Min       1Q    Median        3Q       Max
-0.35474  -0.08862  -0.00121   0.08819   0.35881

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
`(Intercept)`        7.741e+00  8.461e-01   9.149 1.14e-15 ***
fueltypegas         -1.756e-01  6.518e-02  -2.693 0.008020 **
carbodyhardtop      -1.332e-01  1.086e-01  -1.226 0.222381
carbodyhatchback    -2.731e-01  8.395e-02  -3.253 0.001460 **
carbodysedan        -1.682e-01  8.231e-02  -2.044 0.043048 *
carbodywagon        -2.487e-01  8.672e-02  -2.867 0.004842 **
drivewheelrwd        1.586e-01  3.674e-02   4.316 3.16e-05 ***
enginelocationrear   4.864e-01  1.792e-01   2.713 0.007577 **
enginetypel         -1.584e-01  7.356e-02  -2.153 0.033179 *
enginetypeohc        1.151e-01  3.496e-02   3.292 0.001288 **
carwidth             3.406e-02  1.322e-02   2.577 0.011103 *
curbweight           3.211e-04  8.963e-05   3.582 0.000483 ***
horsepower           1.687e-03  7.742e-04   2.179 0.031189 *
citympg             -8.948e-02  1.618e-02  -5.531 1.72e-07 ***
citympg.2            1.274e-03  2.456e-04   5.187 8.15e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1482 on 128 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9998
F-statistic: 3.824e+04 on 15 and 128 DF,  p-value: < 2.2e-16
```
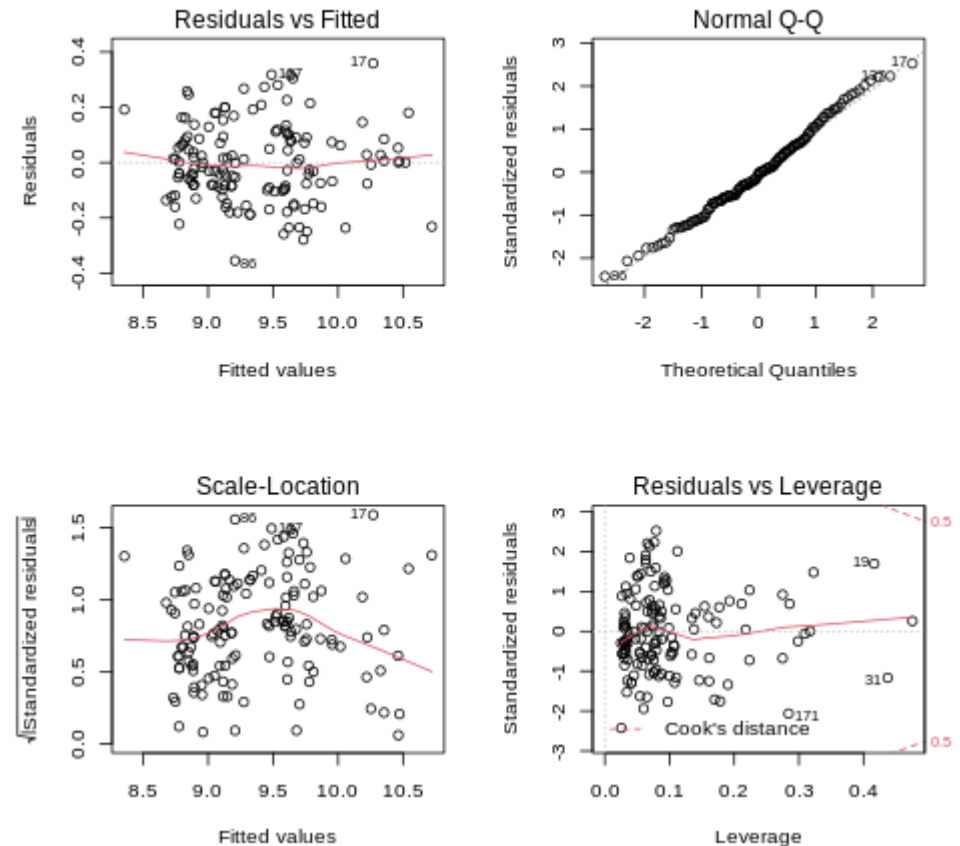

A linear model was fit to the training data with an R-Squared of 0.9998 on the training data and 0.903 on the test data. Along with an RMSE of 0.14 and 0.15 for the training and test data, respectively.

The above plot tests our model assumptions. Overall, the model shows a strong ability to predict the price of a car.