Linear Regression case study

Linear regression Case Study

Problem statement:

This A Chinese automobile company **Geely Auto** aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts is about the pattern recognition of the handwritten digits(0-9) recognition.

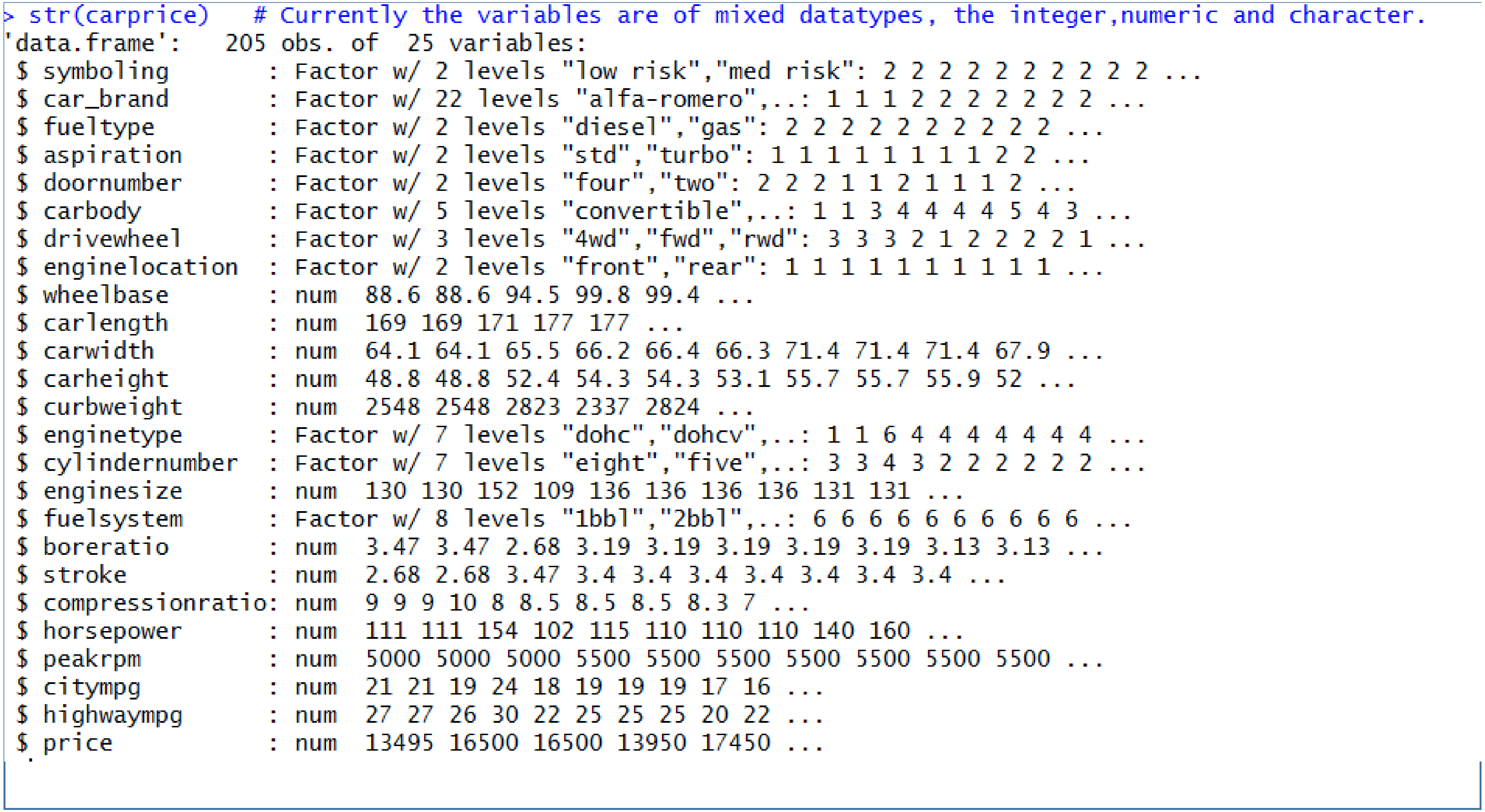
Goal of the case study:

You are required to model the price of cars with the available independent variables.

Problem solving methodology:

1. Importing the test and training datasets
2. Data preparation regarding the Missing values, outliers, blank values, variable transformation(dummy variables creation)
3. EDA (univariate, bivariate, derived metrics)
4. Splitting the data into train and test set
5. Applying stepAIC function
6. Model building process

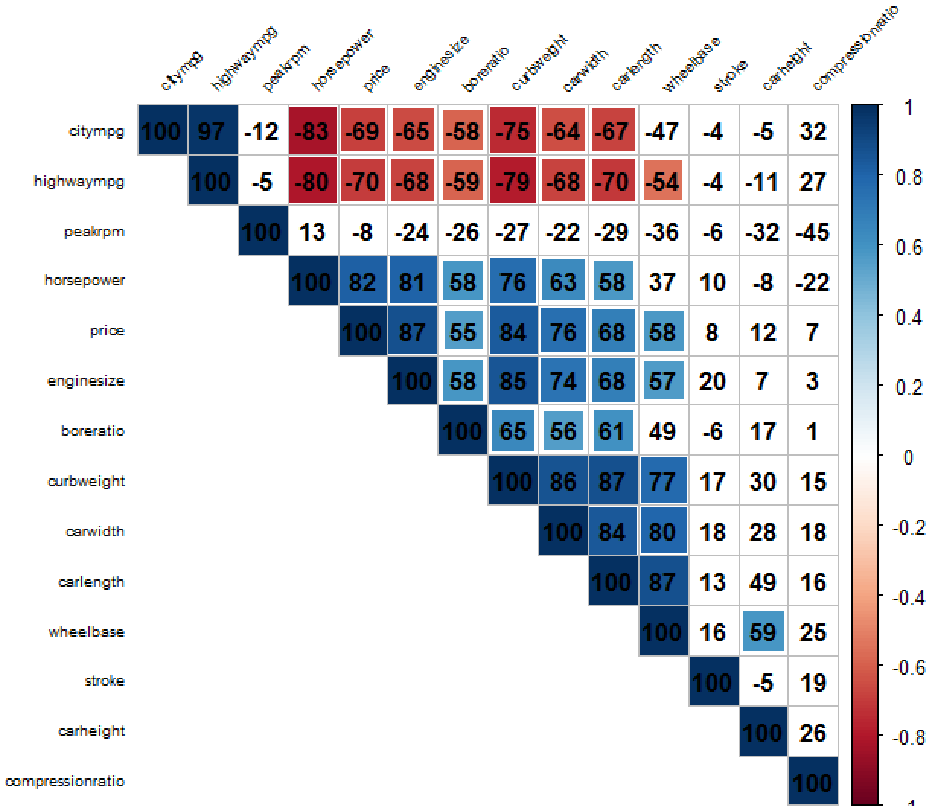
Data Understanding



```
> str(carprice)   # Currently the variables are of mixed datatypes, the integer,numeric and character.
'data.frame':   205 obs. of  25 variables:
 $ symboling       : Factor w/ 2 levels "low risk","med risk": 2 2 2 2 2 2 2 2 2 2 ...
 $ car_brand       : Factor w/ 22 levels "alfa-romero",..: 1 1 1 2 2 2 2 2 2 2 ...
 $ fueltype        : Factor w/ 2 levels "diesel","gas": 2 2 2 2 2 2 2 2 2 2 ...
 $ aspiration      : Factor w/ 2 levels "std","turbo": 1 1 1 1 1 1 1 1 2 2 ...
 $ doornumber      : Factor w/ 2 levels "four","two": 2 2 2 1 1 2 1 1 1 2 ...
 $ carbody         : Factor w/ 5 levels "convertible",..: 1 1 3 4 4 4 4 5 4 3 ...
 $ drivewheel      : Factor w/ 3 levels "4wd","fwd","rwd": 3 3 3 2 1 2 2 2 2 1 ...
 $ enginelocation  : Factor w/ 2 levels "front","rear": 1 1 1 1 1 1 1 1 1 1 ...
 $ wheelbase       : num  88.6 88.6 94.5 99.8 99.4 ...
 $ carlength       : num  169 169 171 177 177 ...
 $ carwidth        : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
 $ carheight       : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
 $ curbweight      : num  2548 2548 2823 2337 2824 ...
 $ enginetype      : Factor w/ 7 levels "dohc","dohcv",..: 1 1 6 4 4 4 4 4 4 4 ...
 $ cylindernumber  : Factor w/ 7 levels "eight","five",..: 3 3 4 3 2 2 2 2 2 2 ...
 $ enginesize      : num  130 130 152 109 136 136 136 136 131 131 ...
 $ fuelsystem      : Factor w/ 8 levels "1bbl","2bbl",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ boreratio       : num  3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13 3.13 ...
 $ stroke          : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4 ...
 $ compressionratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
 $ horsepower      : num  111 111 154 102 115 110 110 110 140 160 ...
 $ peakrpm         : num  5000 5000 5000 5500 5500 5500 5500 5500 5500 5500 ...
 $ citympg         : num  21 21 19 24 18 19 19 19 17 16 ...
 $ highwaympg      : num  27 27 26 30 22 25 25 25 20 22 ...
 $ price           : num  13495 16500 16500 13950 17450 ...
```

EDA findings

1.Check for missing values and blanks 2. Check for Outliers 3. Visualisations



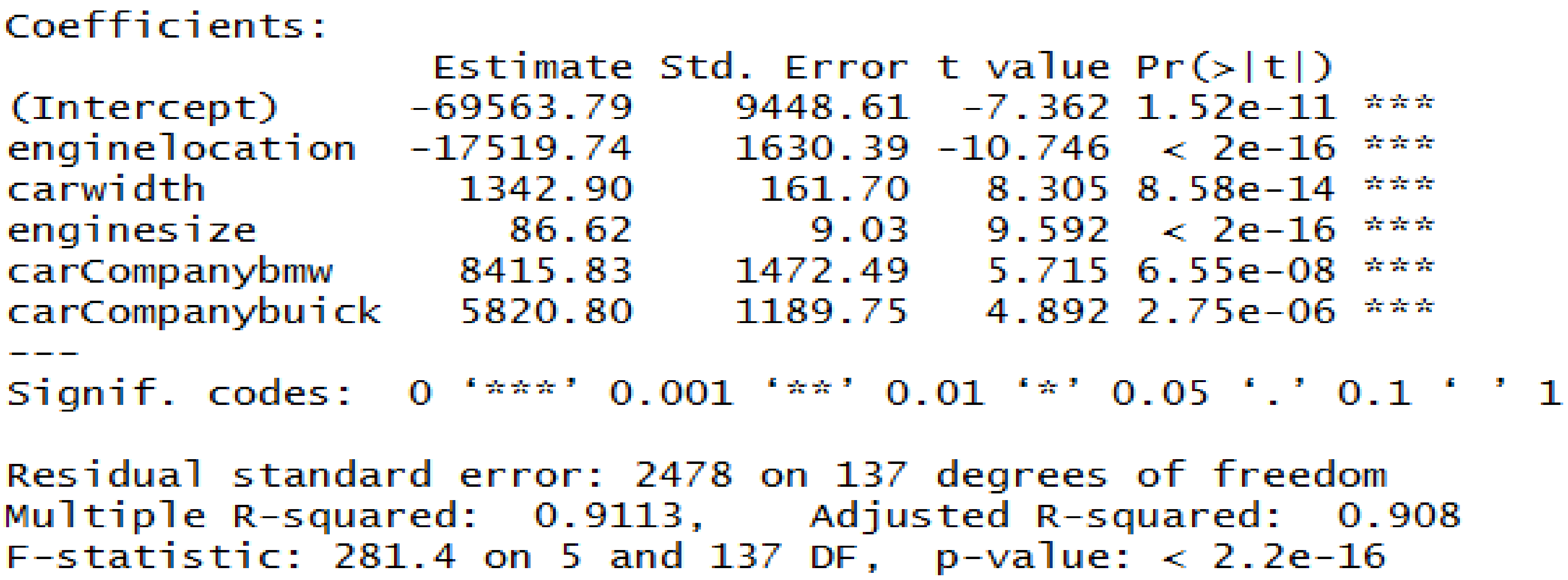
	citympg	highwaympg	peakrpm	horsepower	price	enginesize	boreratio	curbweight	carwidth	carlength	wheelbase	stroke	carheight	compressionratio
citympg	100	97	-12	-83	-69	-65	-58	-75	-64	-67	-47	-4	-5	32
highwaympg		100	-5	-80	-70	-68	-59	-79	-68	-70	-54	-4	-11	27
peakrpm			100	13	-8	-24	-26	-27	-22	-29	-36	-6	-32	-45
horsepower				100	82	81	58	76	63	58	37	10	-8	-22
price					100	87	55	84	76	68	58	8	12	7
enginesize						100	58	85	74	68	57	20	7	3
boreratio							100	65	56	61	49	-6	17	1
curbweight								100	86	87	77	17	30	15
carwidth									100	84	80	18	28	18
carlength										100	87	13	49	16
wheelbase											100	16	59	25
stroke												100	-5	19
carheight													100	26
compressionratio														100

#1. The distribution of the dataset on various attributes
#2. Correlation among the numeric fields and their relationship
#3. Visualisations of the factors impacting the price variable
#4. The curb weight and its impact on the mileage(citympg and highwaympg)
#5. The price points of various car brands
#6. Importance of drive wheels and curb weight

Linear regression concepts

- Linear regression
- Best fit line
- Metrics for assessing strength of linear regression model
  - R2
  - RSE
  - T value and P value
- Dummy variables
- R-squared vs Adjusted R-squared
- Multicollinearity :
- Variable selection methods
- Model evaluation

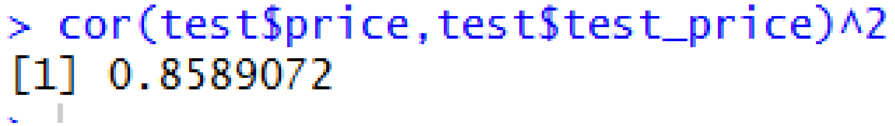
Final Model



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -69563.79    9448.61  -7.362 1.52e-11 ***
enginelocation  -17519.74    1630.39 -10.746  < 2e-16 ***
carwidth          1342.90     161.70   8.305 8.58e-14 ***
enginesize          86.62       9.03   9.592  < 2e-16 ***
carCompanybmw     8415.83    1472.49   5.715 6.55e-08 ***
carCompanybuick   5820.80    1189.75   4.892 2.75e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2478 on 137 degrees of freedom
Multiple R-squared:  0.9113,    Adjusted R-squared:  0.908 
F-statistic: 281.4 on 5 and 137 DF,  p-value: < 2.2e-16
```

Model Evaluation:

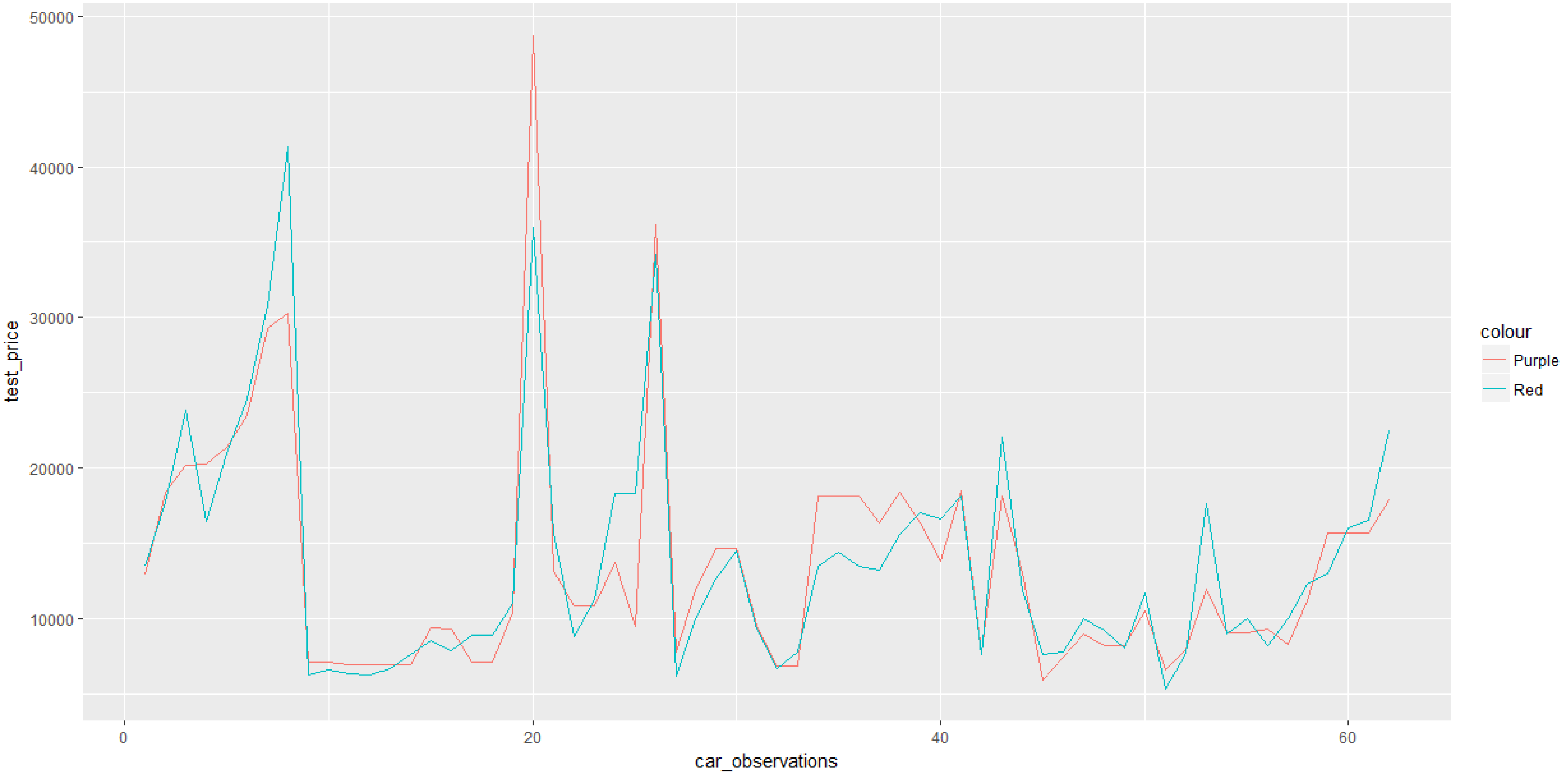


```
> cor(test$price,test$test_price)^2
[1] 0.8589072
```

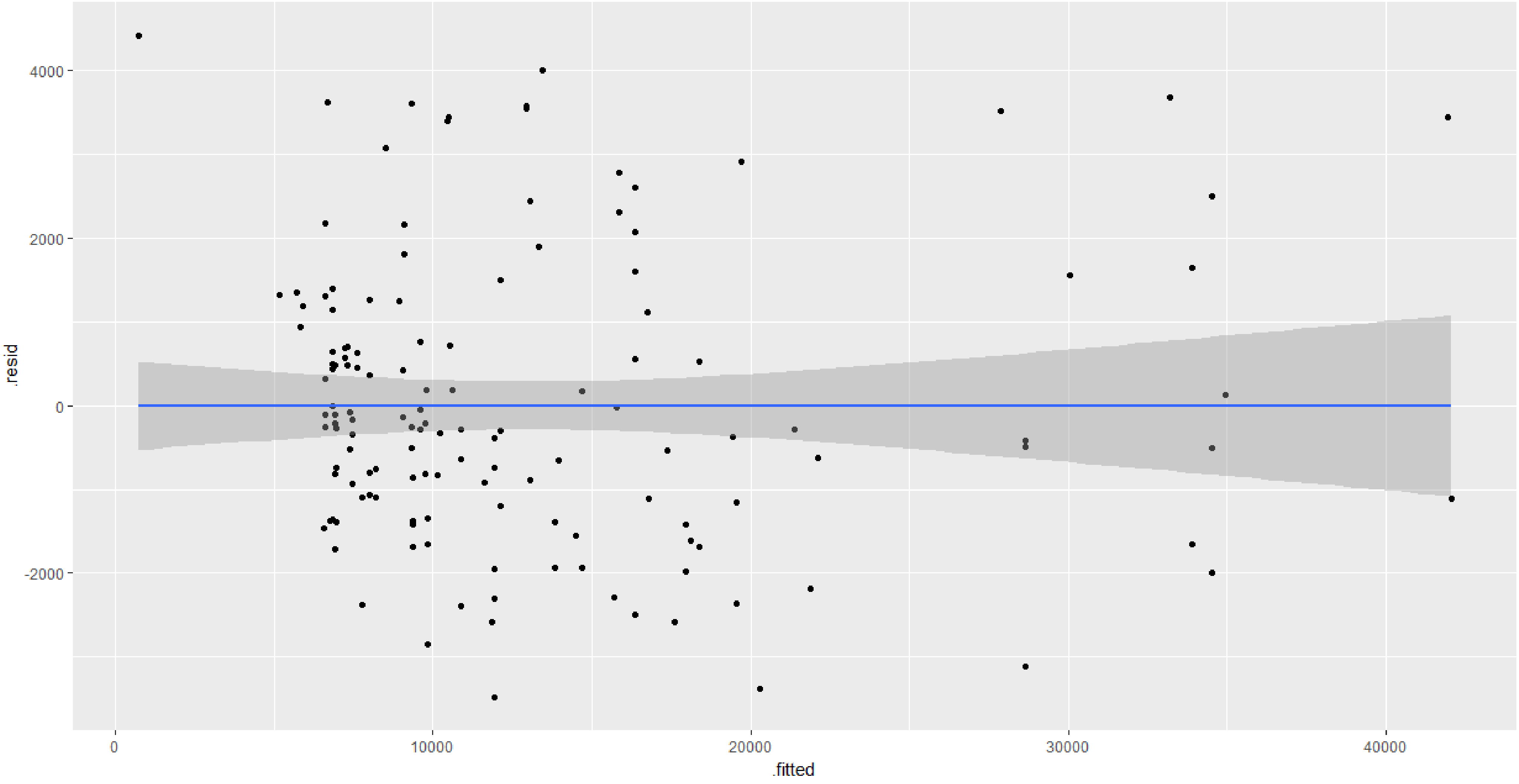
The R2 value in test is 0.85. This is to say the predicted values in the model can explain 85% of variation in the actual outcomes and is fairly good model.

Plotting actual versus predicted values

Actual versus predicted



Plotting residual error


.resid
.fitted