R DATA 컬럼소개 & 모델 선정 방법

- **컬럼소개**

- model : 모델
- year : 연식
- price : 가격
- transmission : 자동, 수동, 세미오토
- mileage : 주행거리
- fuelType : 연료타입 디젤, 휘발유, 기타, 하이브리드, 전기
- tax : 세금
- mpg : 마일퍼 갤런, 연비
- engineSize : 엔진 사이즈 cc

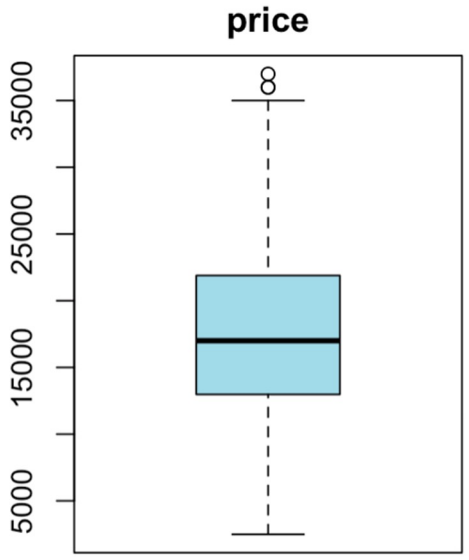
모델 선정

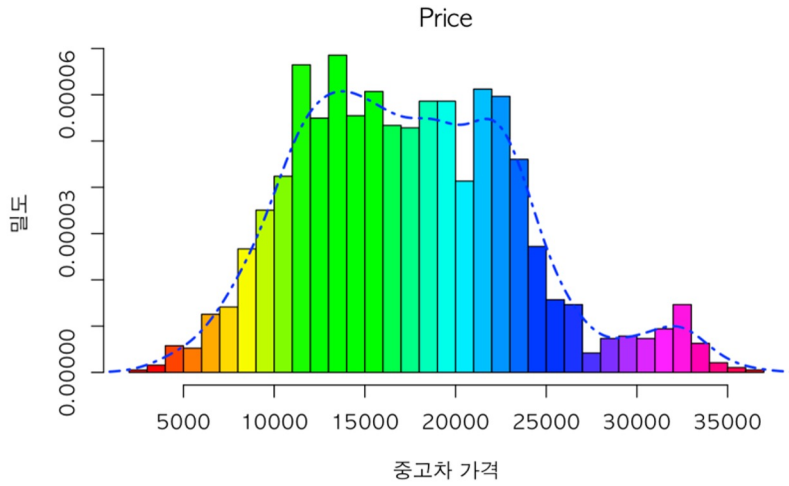
Var1 <fctr>	Freq <int>
A3	1929
Q3	1417
A4	1381
A1	1347
A5	882
Q5	877
Q2	822
A6	748
Q7	397
TT	336
A7	122
A8	118
Q8	69
RS6	39
RS3	33
RS4	31
RS5	29
R8	28
S3	18
SQ5	16
S4	12
SQ7	8
S8	4
S5	3
A2	1

- Data.frame에 등록된 모델 중 Row 수가 가장 많은 모델 3가지를 선정하여 데이터 정제 및 예측모델을 만들기로 결정

Price & Year Column 정제

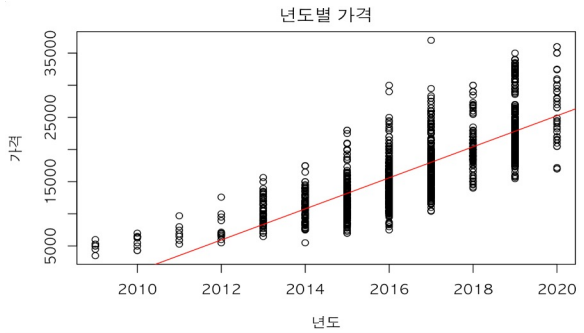
Price Column

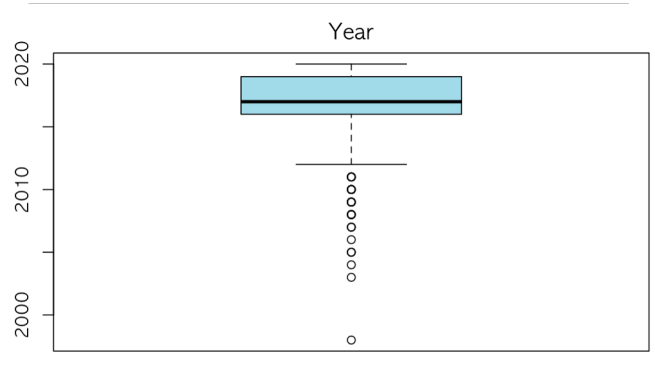




- Target column인 price는 이상치가 확인되지 않고 정규분포를 그리고 있어서 이 모델의 경우 정제를 하지 않았지만, 특이치가 나온 모델은 특이치 제거를 하여 사용함

Year Column

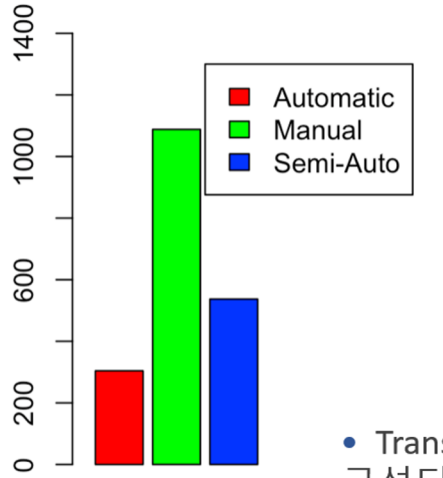


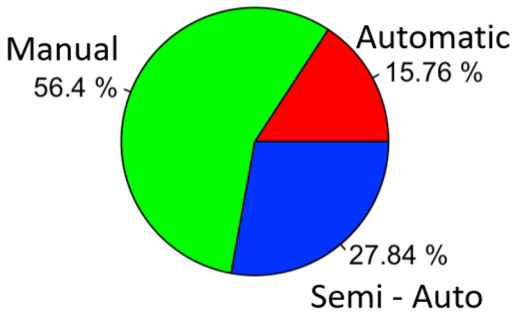


- 근거 :
  - 연도별 가격 데이터를 회귀분석한 결과 회귀선이 시작되기 이전의 year Data는 삭제하기로 결정

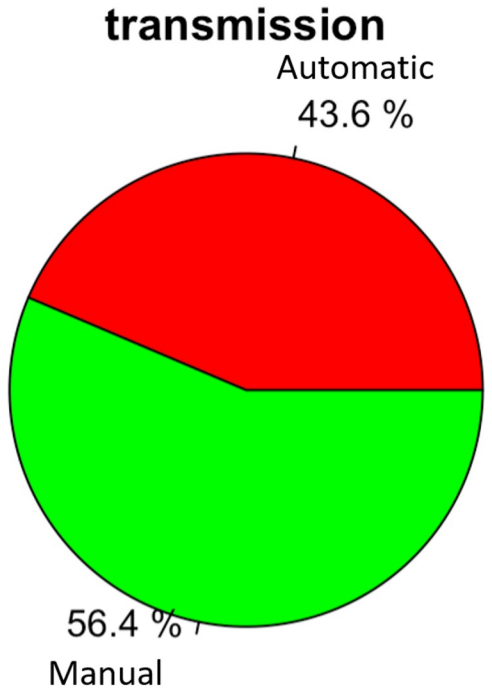
- 결과 : 2010년도 이후 연도부터 사용하기로 결정

Transmission Column 정제





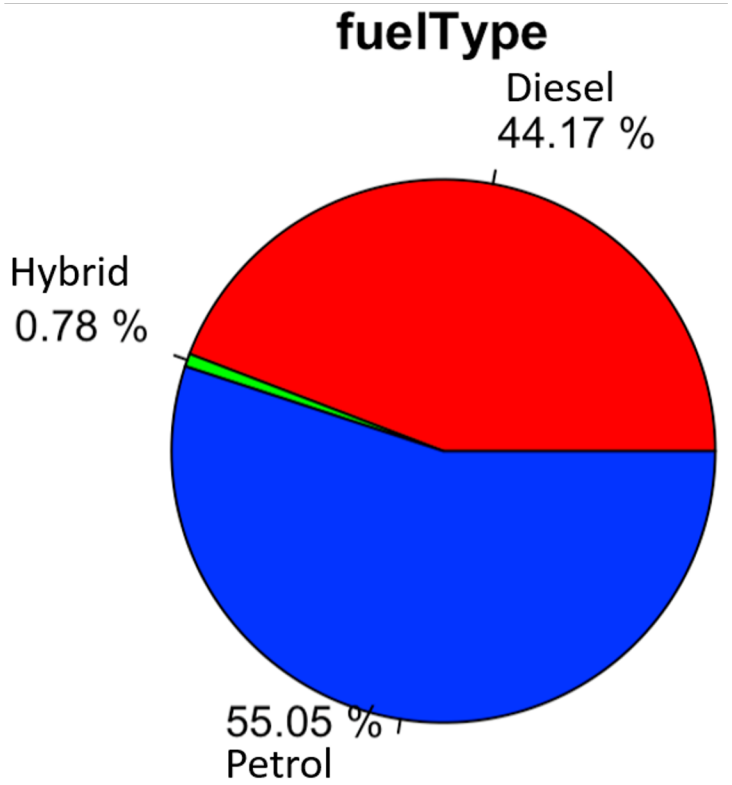
- Transmission Column은 3가지로 구성되어 있고, 셋의 비율이 피처 컬럼으로 사용되기에 적절하다고 판단



- Automatic과 Seme– Auto 컬럼을 합치고, Manual 컬럼과 비교 시 비율의 차이가 있어 피처컬럼으로 사용하기로함.

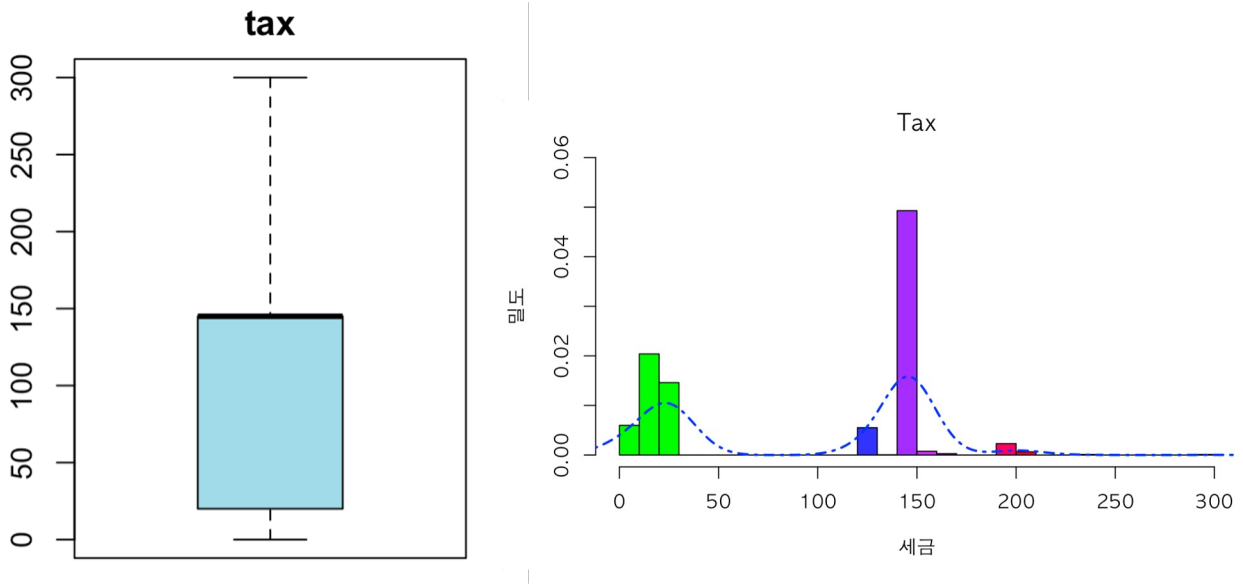
- 합친 이유는 Semi-Auto는 Automatic과의 방식이 거의 일치하며, 클러치가 없는 점, 영국의 중고차사이트(부록 참조)를 통하여 합치게 됨.

Fuel Type & Tax Column 정제



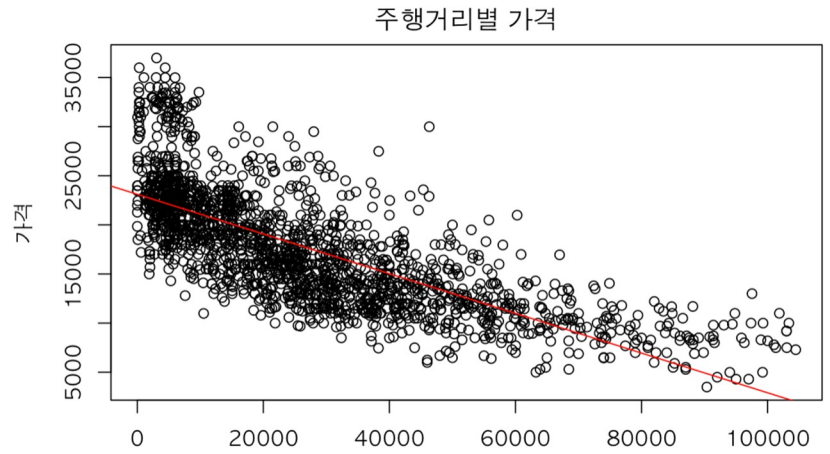
- Fuel Type 컬럼은 Petrol, Diesel, Hybrid, 3가지로 구성되어 있음 Hybrid Column은 0.78%로 데이터가 너무 적어 삭제하고 Petrol과 Diesel Column만을 사용하기로 결정

Tax Column

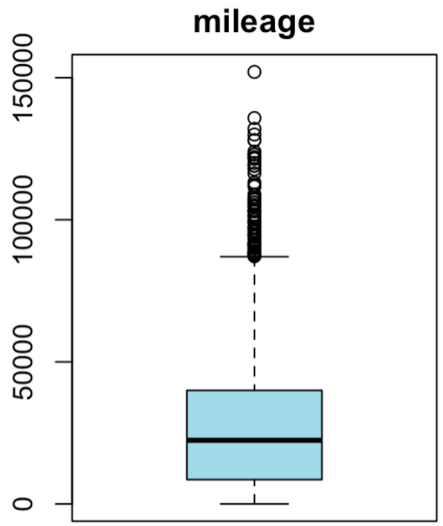


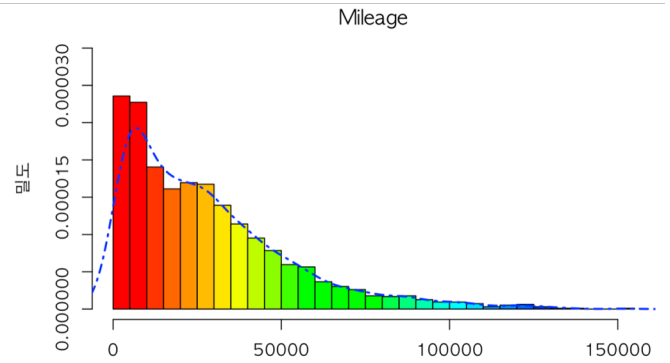
- Tax은 상관계수는 0.66 으로 높은 관계를 나타내지만, 0 값이 132개로 많은 비중을 차지하고 있어, 중위 값이 3사 분위 값은 같아서 제거하기로 결정

Mileage Column 정제

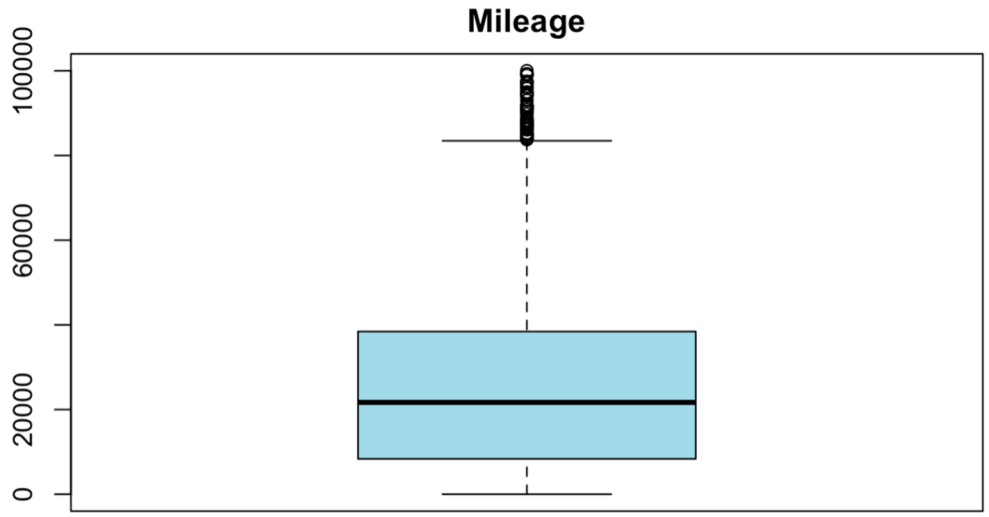


- -0.74 의 역 상관계수를 보여 피처컬럼으로 적합.



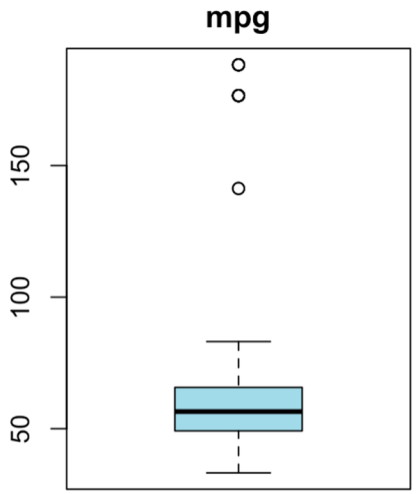


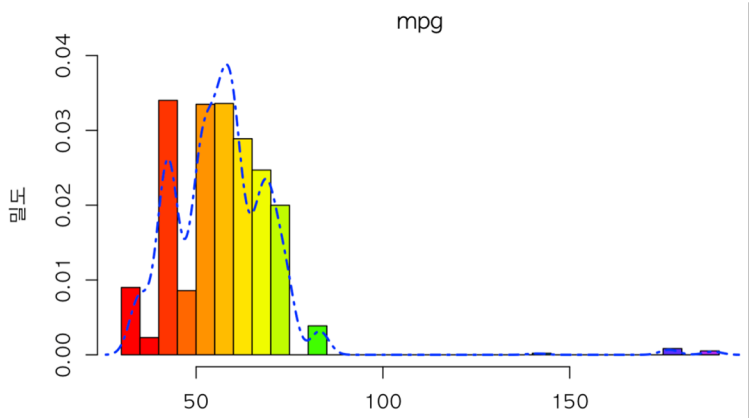
- Histogram과 boxplot에서 Data의 분포도가 일정구간에 집중되어 있지만, 특이치의 값들은 집중된 구간과의 거리가 먼 것을 알 수 있고, 데이터가 없다는 것을 알 수 있음. 이러한 결과로 특이치를 확인하고, 이를 제거하여 사용하기로 결정함



- 정제가 완료된 mileage Column

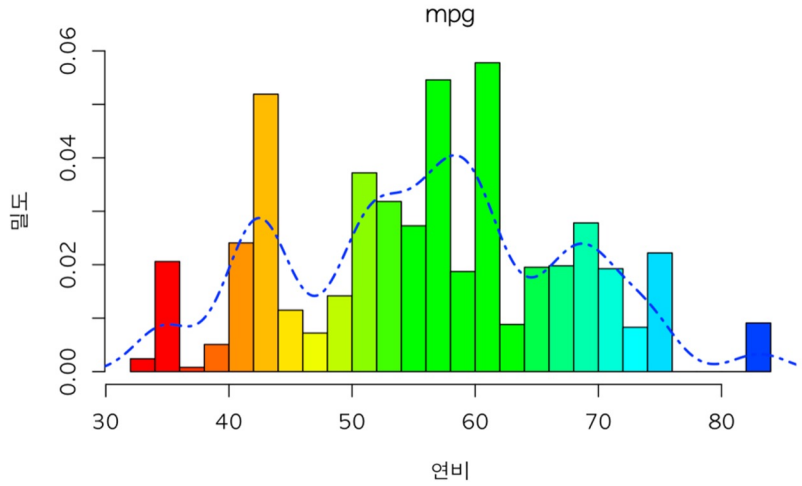
Mpg Column 정제

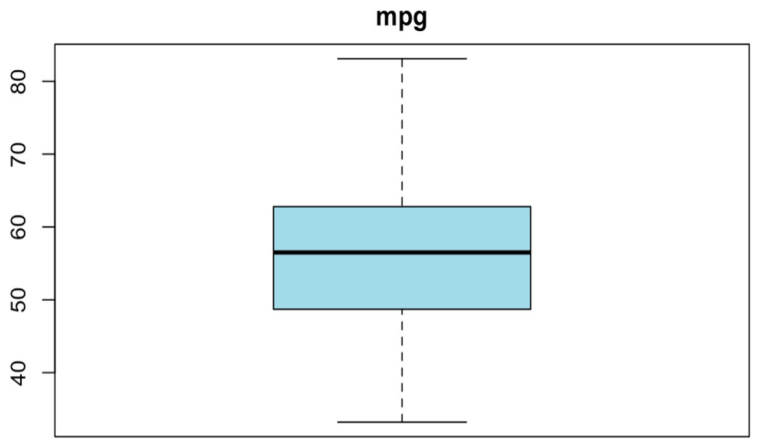




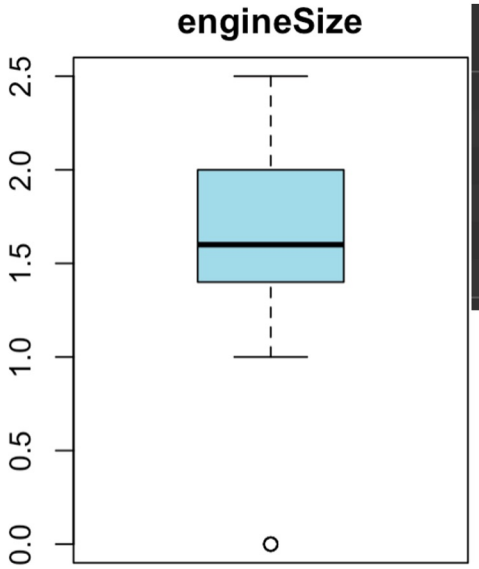
- Histogram과 boxplot에서 Data의 분포도가 일정구간에 집중되어 있지만, 특이치의 값들은 집중된 구간과의 거리가 먼 것을 알 수 있고, 데이터가 없다는 것을 알 수 있음. 이러한 결과로 특이치를 확인하고, 이를 제거하여 사용하기로 결정함

- 정제가 완료된 mpg Histogram과 boxplot





Engine Size 정제



model <chr>	fuelType <chr>	engineSize <dbl>
A3	Petrol	0
A3	Petrol	0
A3	Petrol	0
A3	Petrol	0
A3	Petrol	0
A3	Petrol	0

- EngineSize 값 0인 것을 확인해본 결과 연료타입이 Petrol 인 것을 확인했다. 배기량이 0값이 될 수가 없으므로 제거하기로 결정

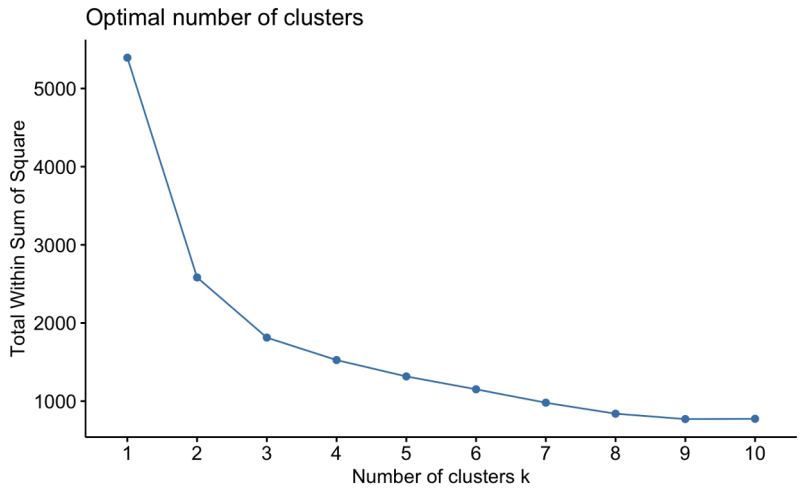
```
Call:
lm(formula = price ~ ., data = Audi_engine)

Residuals:
    Min      1Q  Median      3Q     Max
-8093.0 -1463.7   -87.2  1386.9 13262.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.535e+06  8.512e+04  -29.78   <2e-16 ***
year         1.269e+03  4.214e+01   30.11   <2e-16 ***
mileage     -6.119e-02  3.894e-03  -15.71   <2e-16 ***
mpg         -2.246e+02  5.865e+00  -38.30   <2e-16 ***
engineSize   4.449e+03  1.794e+02   24.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2278 on 1864 degrees of freedom
Multiple R-squared:  0.8564,    Adjusted R-squared:  0.8561
F-statistic:  2779 on 4 and 1864 DF,  p-value: < 2.2e-16
```

	price	year	mileage	mpg	engineSize
price	1.0000000	0.78624885	-0.7410043	-0.77060182	0.20070897
year	0.7862488	1.00000000	-0.7593443	-0.51858456	-0.05861291
mileage	-0.7410043	-0.75934429	1.0000000	0.55372440	0.12343851
mpg	-0.7706018	-0.51858456	0.5537244	1.00000000	-0.07215874
engineSize	0.2007090	-0.05861291	0.1234385	-0.07215874	1.00000000

- Engine Size 상관계수는 0.2로 관련이 없다고 나오지만, 다중회귀분석 결과 연관성이 ‘***’로 높다고 나와, 사용하기로 결정

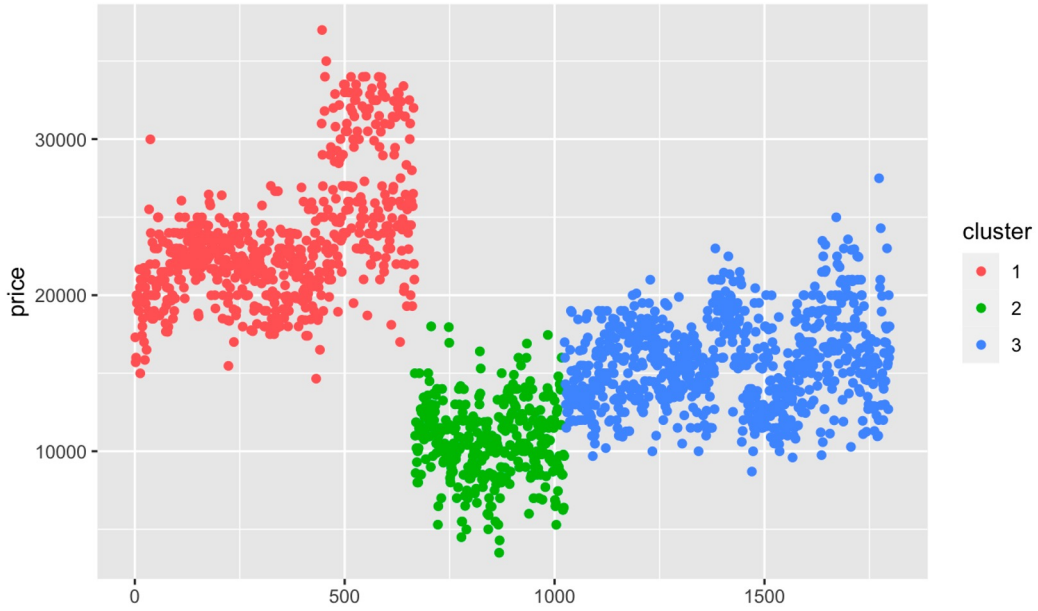
K means & 군집분포

K means



- library(factoextra) 안에 있는 fviz_nbclust 함수를 사용하여 Elbow 값을 확인하여 Kmeans center값으로 사용.

군집분포



- 클러스터를 분포도로 표현하여 군집이 잘 형성되어있는 것을 확인하여 kmeans가 잘 되었는지 확인을 한 후 target column인 price의 범위로 지정하여 사용

브랜드 모델 별 예측모델 표

모델이름/예측모델	Random Forest	SVM	인공신경망	의사결정나무
A - Class	0.9684639	**0.9777213**	0.2875339	0.960529
C - Class	0.9684639	**0.9877213**	0.4922686	0.960529
E - Class	0.9511737	**0.9816887**	0.4286385	0.6044601
1 Series	0.9469428	**0.9605523**	0.2706114	0.9364892
3 Series	0.9142066	**0.954428**	0.3404059	0.9396679
5 Series	0.9365741	**0.95**	0.3166667	0.9097222
A3	**0.9632432**	0.9565766	0.2926126	0.9302703
A4	**0.99631**	0.9745387	0.5121771	0.9704797
Q3	0.9587121	**0.9643939**	0.3049242	0.9291667
Golf	**0.9755906**	0.9750895	0.3717251	0.9605583
Polo	0.9898075	**0.9908267**	0.3761042	0.9833522
Tiguan	**0.9762745**	0.9676471	0.3472549	0.9498039
Fiesta	**0.9916471**	0.9807059	0.3632941	0.9828235
Focus	0.9809015	**0.9851031**	0.3748663	0.9787624
Kuga	0.9695853	**0.9794163**	0.2666667	0.9466974