중고차 가격 예측모델 개발 및 모델 성능 향상방안

청년 AI·빅데이터 아카데미 21기

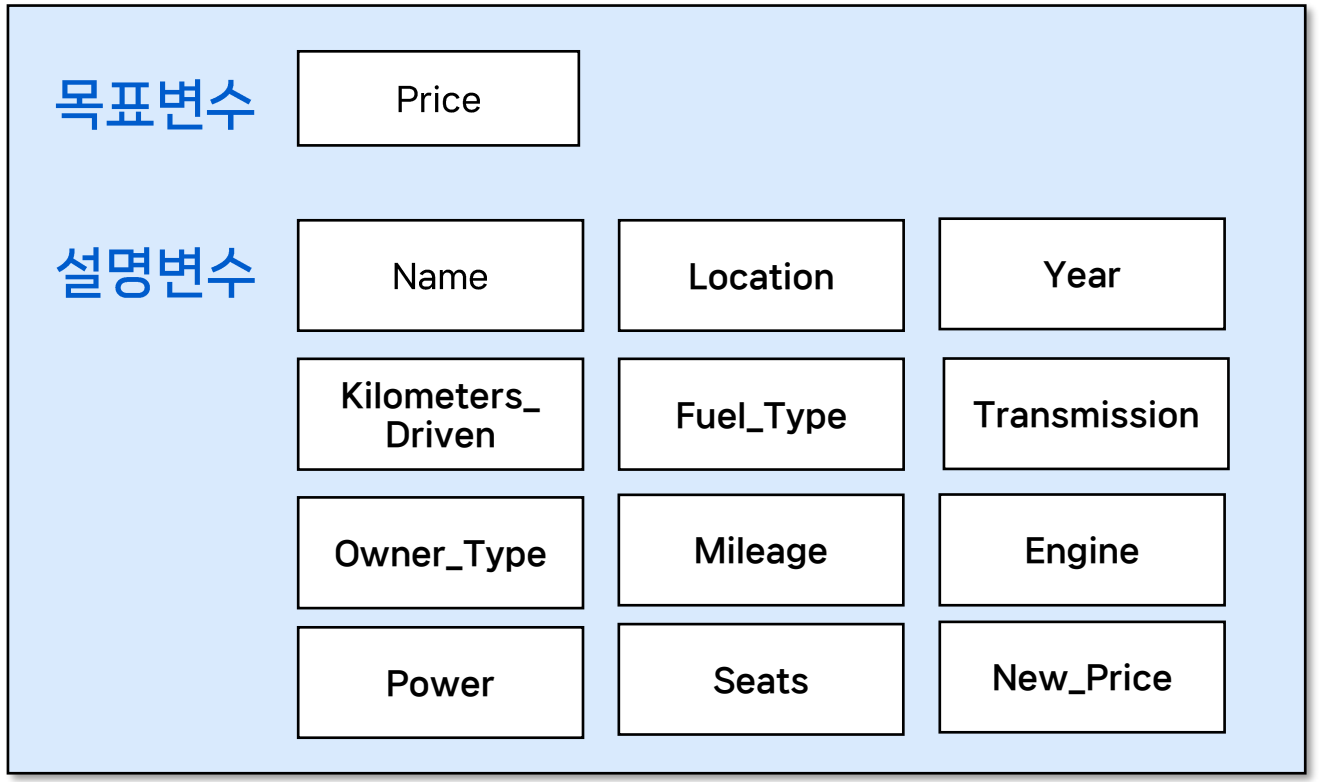
BigData 종합실습1
B4 권혁준

Contents

인도에서는 최근 신차 판매가 주춤하면서, 중고차 시장은 지속적인 성장세를 보이고 있어 신차 시장보다 규모가 더 커질 전망이다. POS_Car㈜는 인도의 중고차 시장에 뛰어들어 신규사업을 목표로 하는 신생 스타트업이다. 신차와 달리 중고차는 가격과 공급 모두에서 엄청난 불확실성을 가진 매물이다. 특히, 고객들은 차량의 특성과 상태(Brand, 차량 연식, 주행거리 등)가 좋으면서, 가격은 저렴한 자동차. 즉, 가성비가 좋은 차를 매우 선호한다.

그러므로 POS_Car㈜는 인도의 중고 자동차 시장에 진출하여 경쟁력 확보와 수익성 향상을 위하여 중고차 가격을 효과적으로 예측할 수 있는 "**핵심영향인자 도출과 가격예측모델** " 을 개발하고자 한다.

변수	변수설명
Price	중고차 가격 중고차 가격 (단위: 천원)
Name	자동차의 브랜드와 모델
Location	자동차를 팔거나 구매할 수 있는 위치
Year	모델의 년도 혹은 버전
Kilometers_ Driven	이전 소유주의 차량 주행거리(Km)
Fuel_Type	자동차의 사용 연료의 종류
Transmission	자동차의 사용 변속기의 종류
Owner_Type	소유권이 직접 소유인지, 중고 소유인지 여부
Mileage	자동차 회사가 제공하는 표준 주행거리(kmpl)
Engine	엔진의 배기량(cc)
Power	엔진의 최대 출력(bhp)
Seats	차의 좌석 수
New_Price	뉴모델의 가격



목표변수인 Price의 핵심영향인자 도출을 통해
성능이 높은 가격예측모델을 개발한다.

- 범주형, 연속형 항목의 특성값 확인

	Name	Location	Price	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price
0	Maruti Wagon R LXI CNG	Mumbai	2682.68	2010	72000	CNG	Manual	First	26.6 kmpl	998 CC	58.16 bhp	5.0	NaN
1	Hyundai Creta 1.6 CRDi SX Option	Pune	19162.00	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	NaN
2	Honda Jazz V	Chennai	6898.32	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.61 Lakh
3	Maruti Ertiga VDI	Chennai	9197.76	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	27194.71	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	NaN

- 범주형, 연속형 항목의 특성값 처리

1) 단위가 들어있는 컬럼의 단위 제외
2) 범주형 항목을 연속형 항목으로 변경(replace(float))
3) Name컬럼을 Brand 컬럼으로 변환 (파생변수 생성)

	Location	Price	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Brand
0	Mumbai	2682.68	2010	72000	CNG	Manual	First	26.60	998.0	58.16	5.0	NaN	Maruti
1	Pune	19162.00	2015	41000	Diesel	Manual	First	19.67	1582.0	126.20	5.0	NaN	Hyundai
2	Chennai	6898.32	2011	46000	Petrol	Manual	First	18.20	1199.0	88.70	5.0	8.61 Lakh	Honda
3	Chennai	9197.76	2012	87000	Diesel	Manual	First	20.77	1248.0	88.76	7.0	NaN	Maruti
4	Coimbatore	27194.71	2013	40670	Diesel	Automatic	Second	15.20	1968.0	140.80	5.0	NaN	Audi

- 범주형, 연속형 항목의 결측치 확인

Location	0
Price	1053
Year	0
Kilometers_Driven	0
Fuel_Type	0
Transmission	0
Owner_Type	0
Mileage	2
Engine	46
Power	46
Seats	53
New_Price	6247
Brand	0

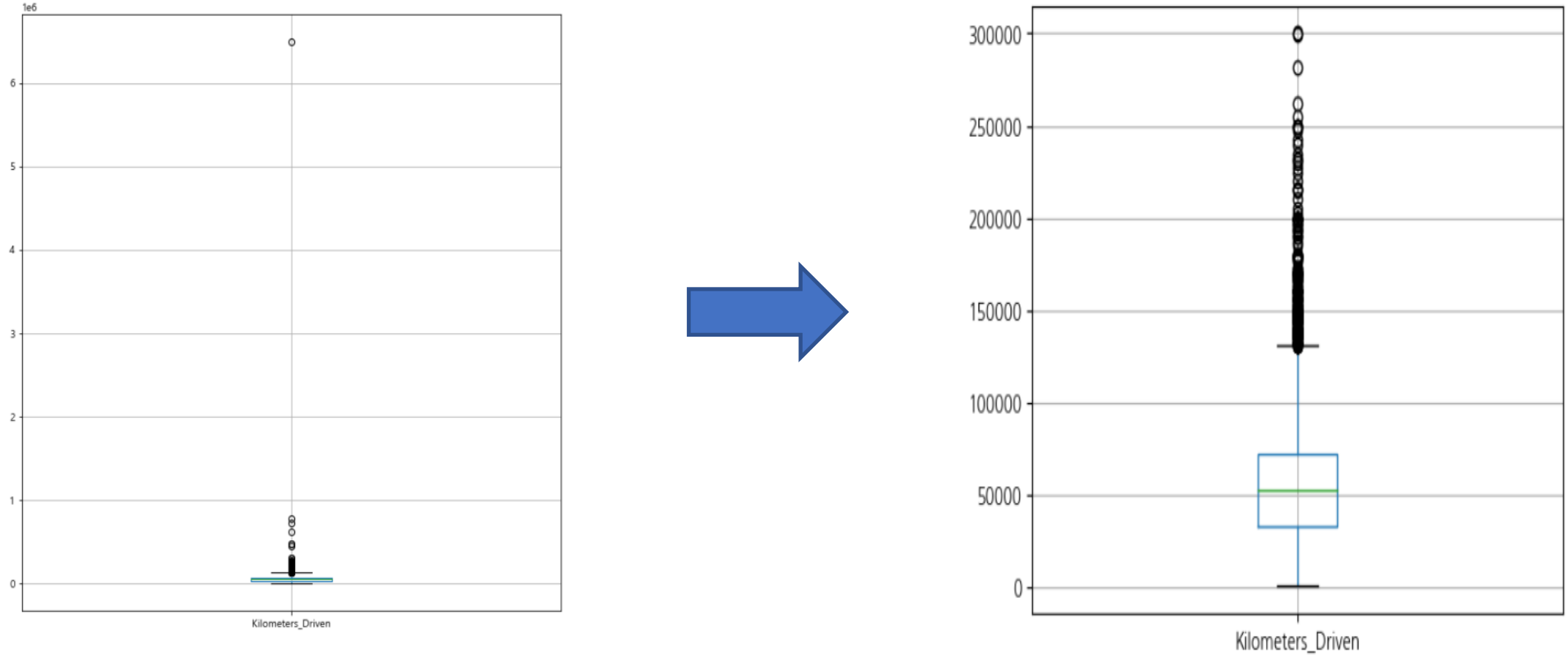
- 범주형, 연속형 항목의 결측치 처리

1) 연속형 설명변수의 결측치: 평균값
2) 연속형 목표변수(Price)항목의 결측치는 제외
3) New_Price 컬럼 제외

Location	0
Price	0
Year	0
Kilometers_Driven	0
Fuel_Type	0
Transmission	0
Owner_Type	0
Mileage	0
Engine	0
Power	0
Seats	0
Brand	0

- 범주형, 연속형 항목의 이상치 확인 및 처리

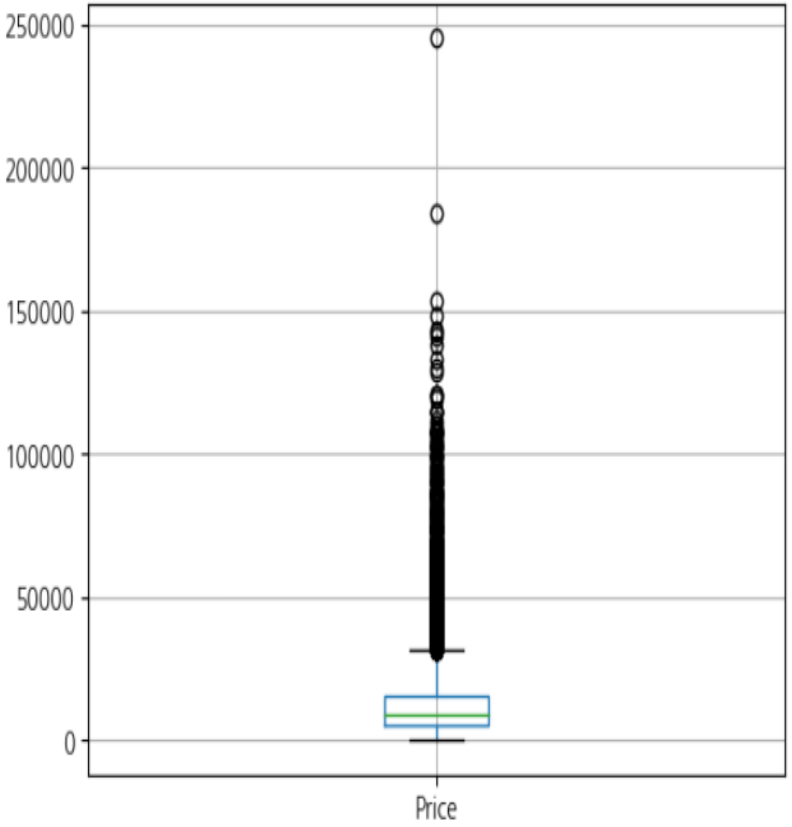
650km는 나올 수 없는 주행거리라 판단하여 이상치로 판단 후, 제거
평균 자동차 주행거리는 2~30만km로 판단, 40만km이상은 이상치로 판단 후 제거

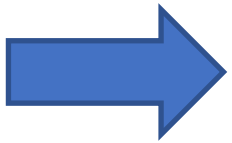


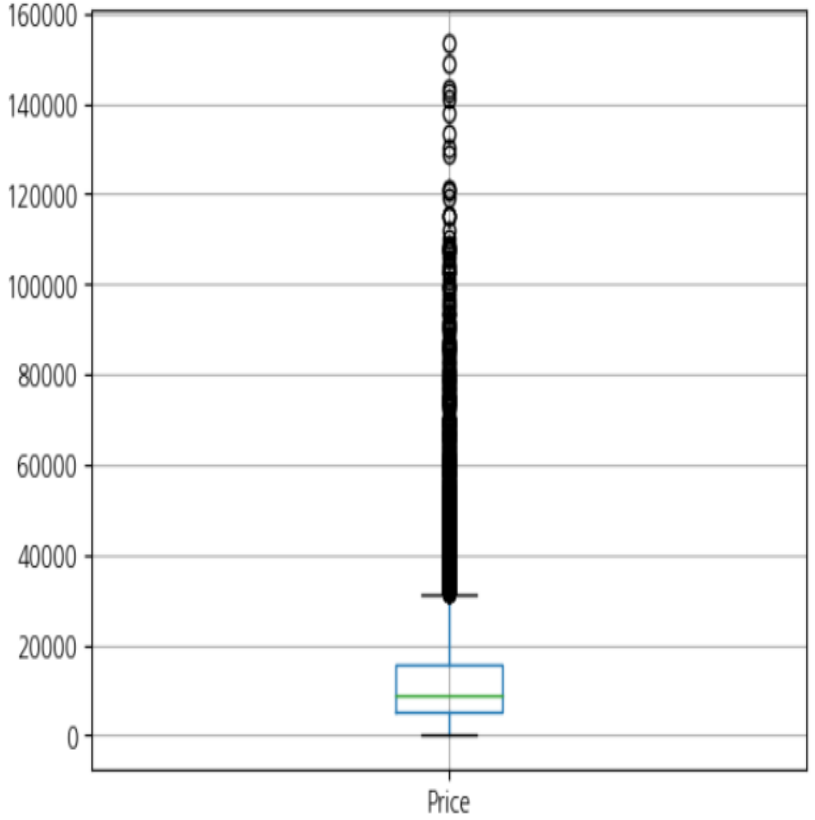
Kilometers_ Driven 이상치 확인 및 처리

- 범주형, 연속형 항목의 이상치 확인

레인지로버의 경우, 신차 가격이 약 2억 500만원으로 중고가격이 2억 4천만원은 이상치로 판단으로 제거
람보르기니의 경우, 신차 가격이 3-4억으로 중고가격 1억 8천만원은 이상치로 판단 안함.
또한, 100보다 작은 값은 이상치로 판단.



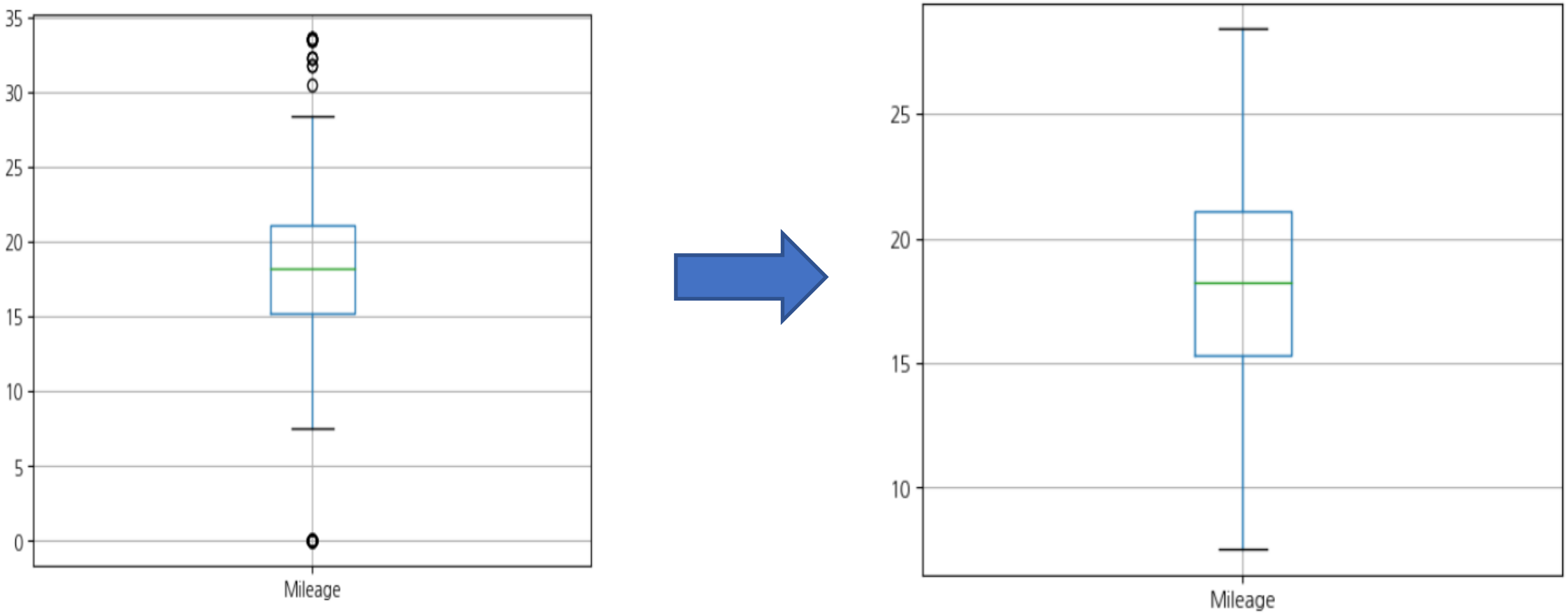




Price 이상치 확인 및 처리

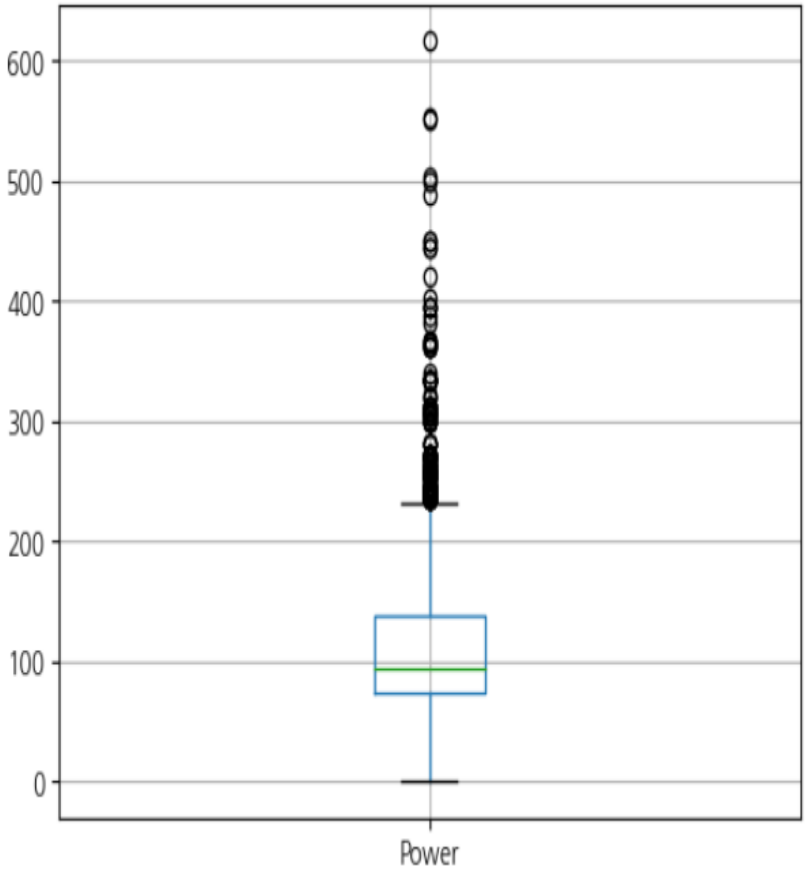
- 범주형, 연속형 항목의 이상치 확인

Mileage가 0은 차가 작동하지 않는 상태라 판단하여, 이상치 제거
자동차 평균 연비를 고려하여 30 이상인 Mileage 이상치 제거



Mileage 이상치 확인 및 처리

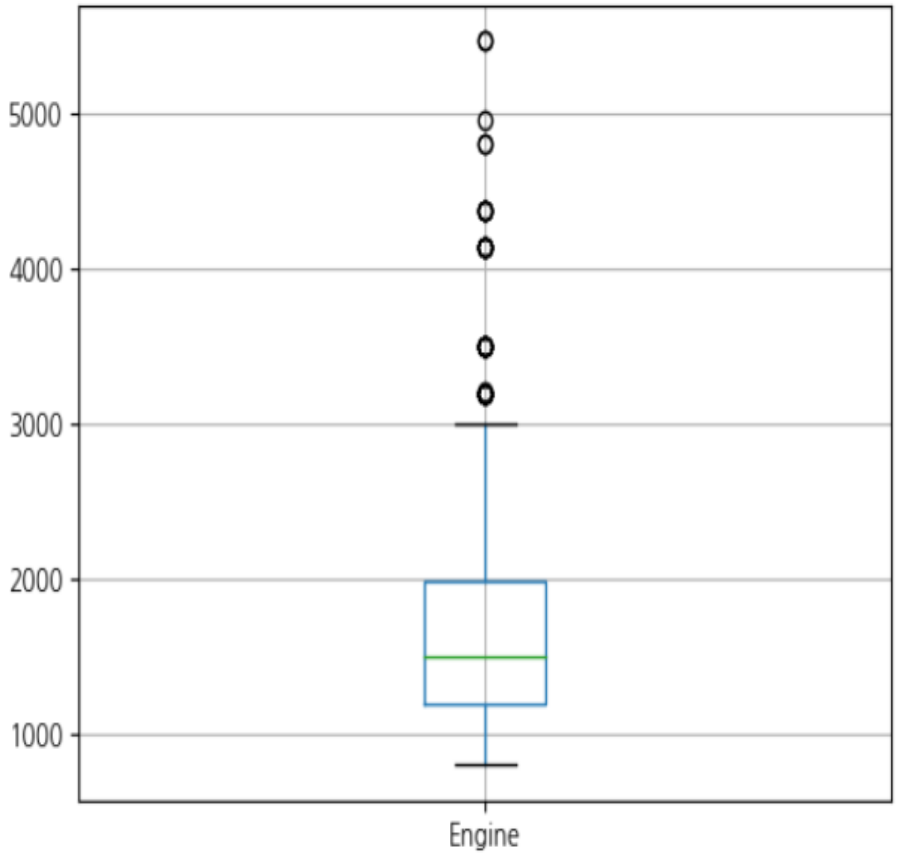
- 범주형, 연속형 항목의 이상치 확인



Power 이상치 확인

평균 마력은 골고루 분포되어 있기 때문에,
이상치라 판단하지 않아 제거하지 않았음

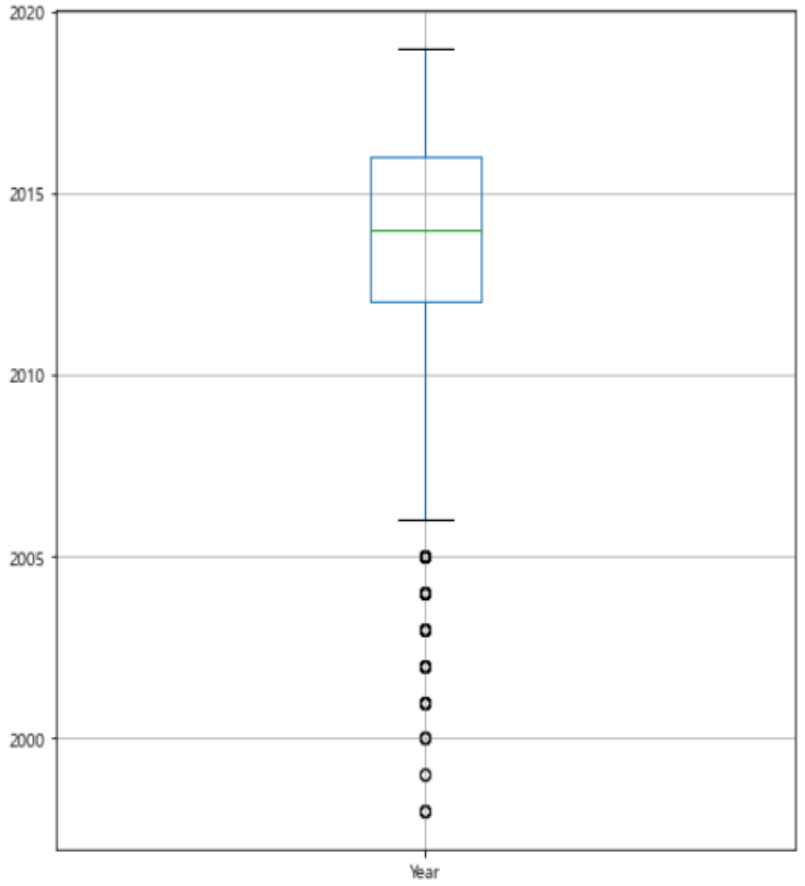
- 범주형, 연속형 항목의 이상치 확인



Engine 이상치 확인

엔진의 최대 출력량은 골고루 분포되어 있기 때문에,
이상치라 판단하지 않아 제거하지 않았음.

- 범주형, 연속형 항목의 이상치 확인

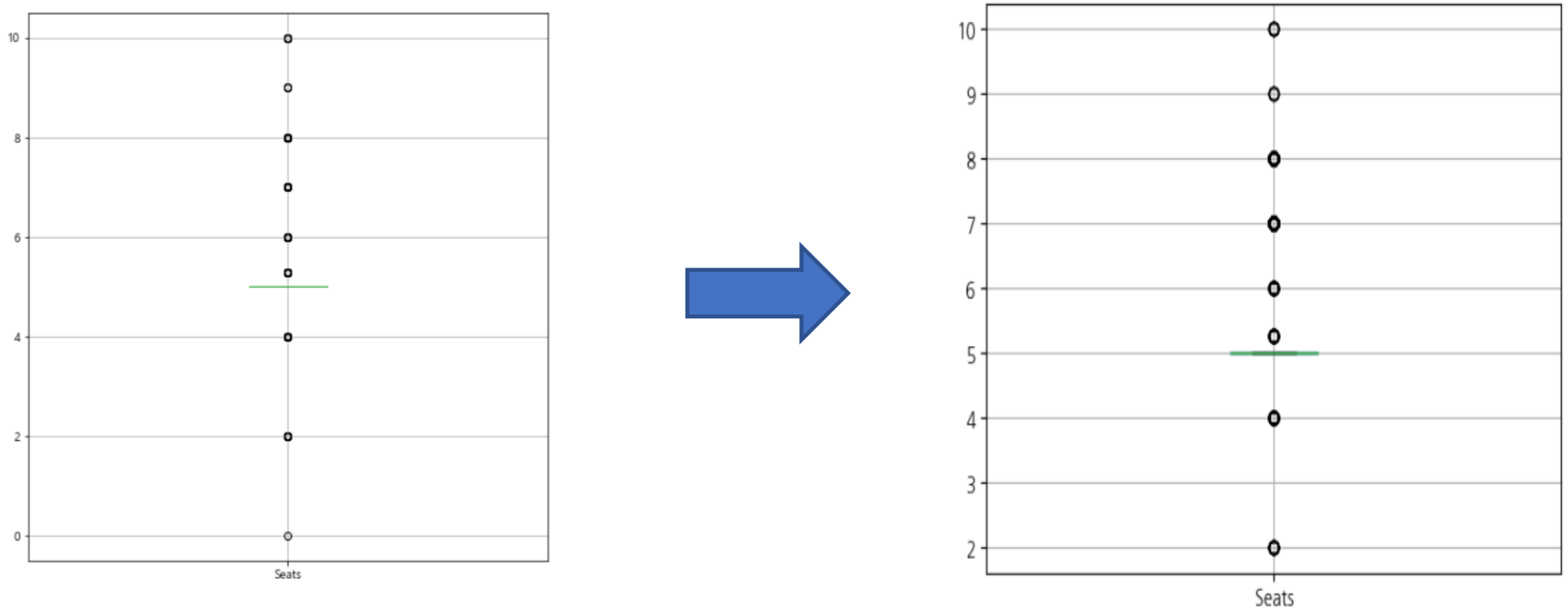


Year 이상치 확인

YEAR은 골고루 분포되어 있기 때문에,
이상치라 판단하지 않아 제거하지 않았음.

- 범주형, 연속형 항목의 이상치 확인

Seats 개수가 0인 행 삭제



Seats 이상치 확인 및 처리

- 파생변수 생성

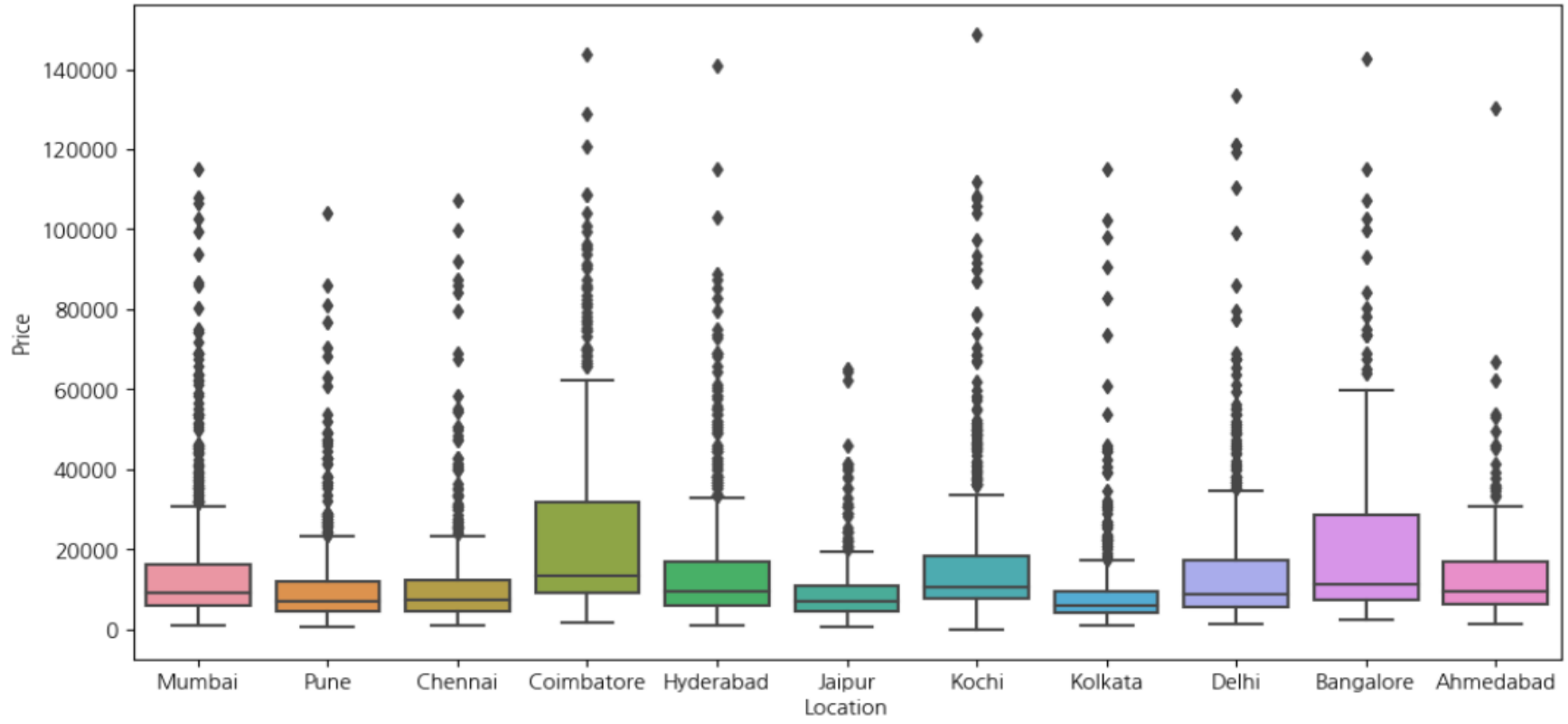
- 단위 수 조절을 위해, 생산 년도 대신 연식으로 바꿈
-> df['Year'] = 2023 - df['Year']

- 요약통계량 확인

	Price	Year	Kilometers_Driven	Mileage	Engine	Power	Seats
count	5857.000000	5857.000000	5857.000000	5857.000000	5857.000000	5857.000000	5857.000000
mean	15150.965667	9.425303	56560.961926	18.308842	1638.621931	114.348541	5.300537
std	17143.427324	3.141733	33600.267336	4.105661	569.736995	50.357096	0.798949
min	689.830000	4.000000	600.000000	7.500000	799.000000	52.800000	2.000000
25%	5748.600000	7.000000	33000.000000	15.290000	1198.000000	78.900000	5.000000
50%	9044.460000	9.000000	52191.000000	18.200000	1497.000000	98.600000	5.000000
75%	16479.320000	11.000000	72022.000000	21.100000	1991.000000	140.000000	5.000000
max	148804.430000	25.000000	300000.000000	28.400000	5461.000000	395.000000	10.000000

- 범주형 설명변수 항목과 목표변수 간의 Boxplot

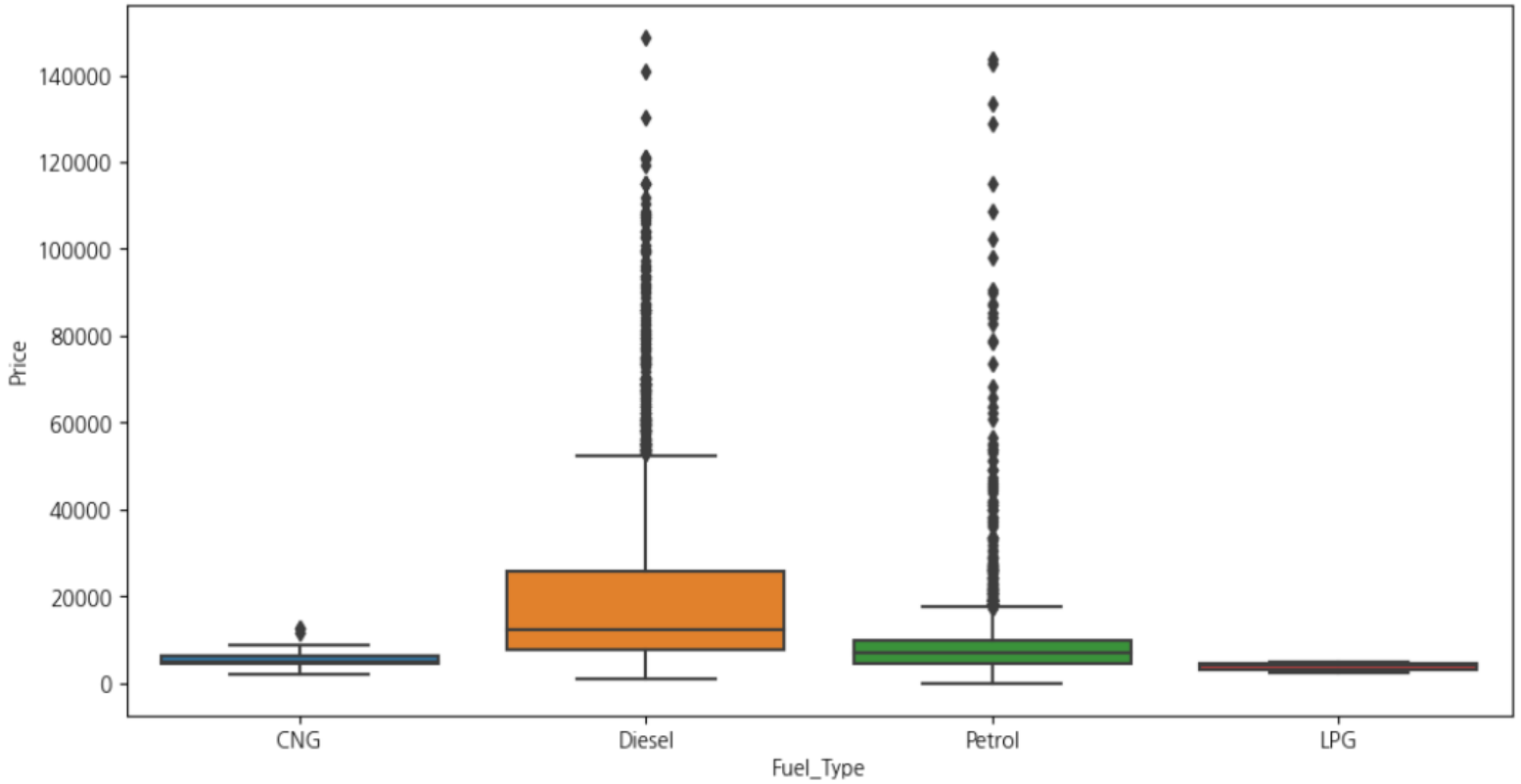
Location과 Price의 분포



boxplot을 보면 차를 팔거나 구매하는 위치에서 뚜렷한 관계성이 보이지 않는다고 생각한다.

- 범주형 설명변수 항목과 목표변수 간의 Boxplot

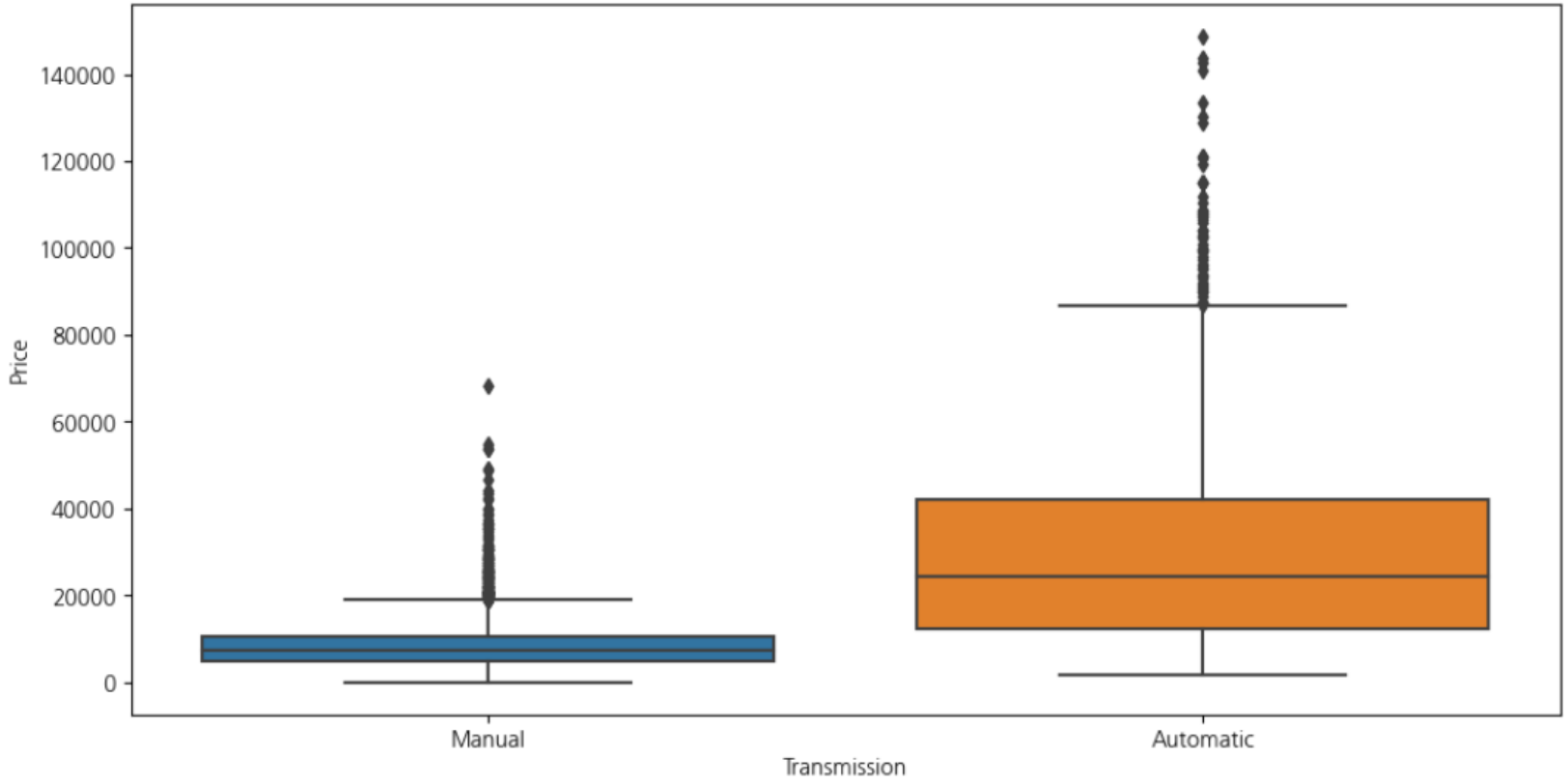
Fuel_Type과 Price의 분포



사용연료에 따른 가격 차이를 보면, Diesel, Petrol이 높은 가격에 형성되어 있는 것을 볼 수 있다.

- 범주형 설명변수 항목과 목표변수 간의 Boxplot

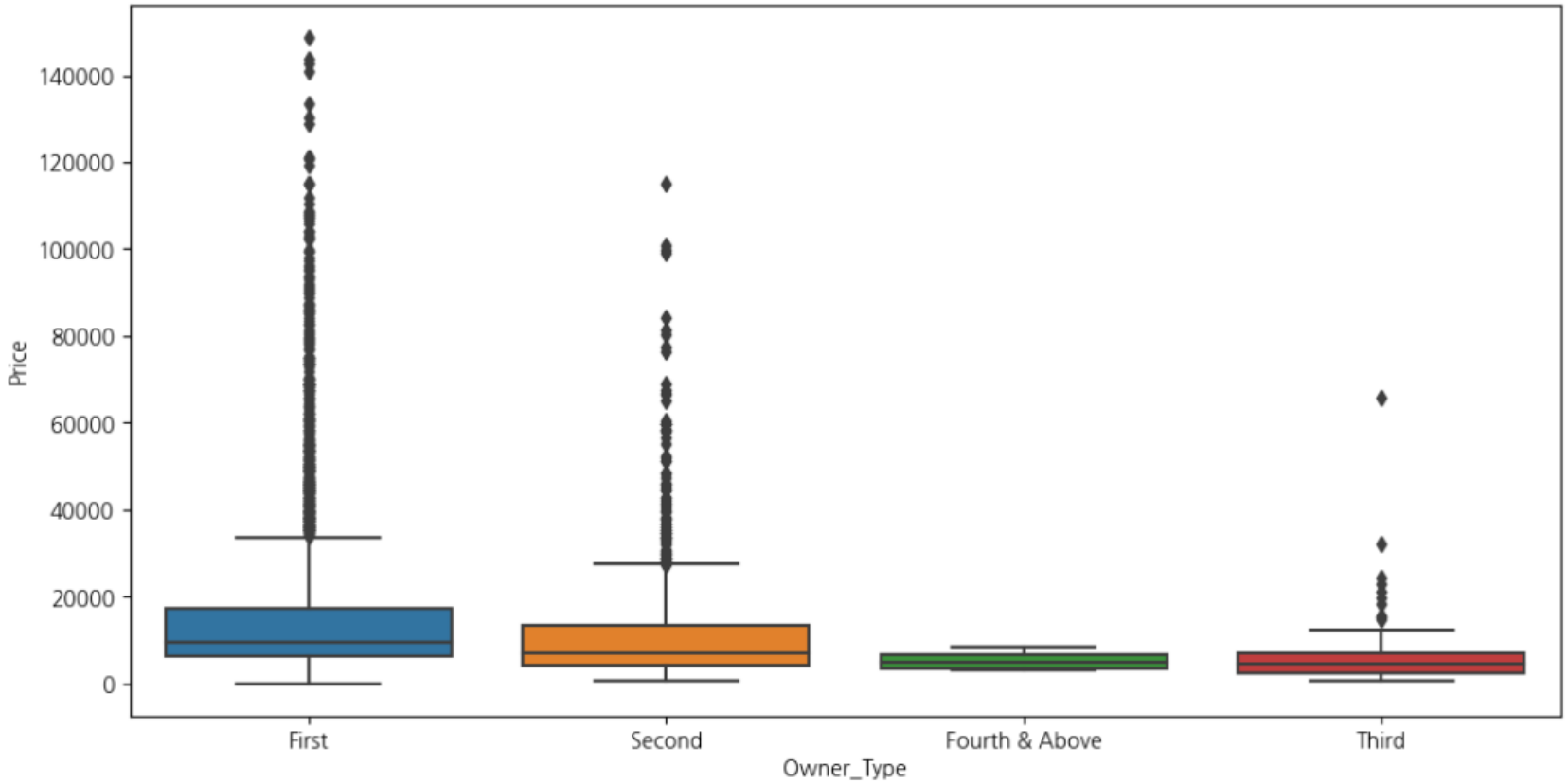
Transmission과 Price의 분포



변속기의 종류에 따른 가격을 보면 수동보다 자동일 경우 높은 가격이 형성되어 있다.

- 범주형 설명변수 항목과 목표변수 간의 Boxplot

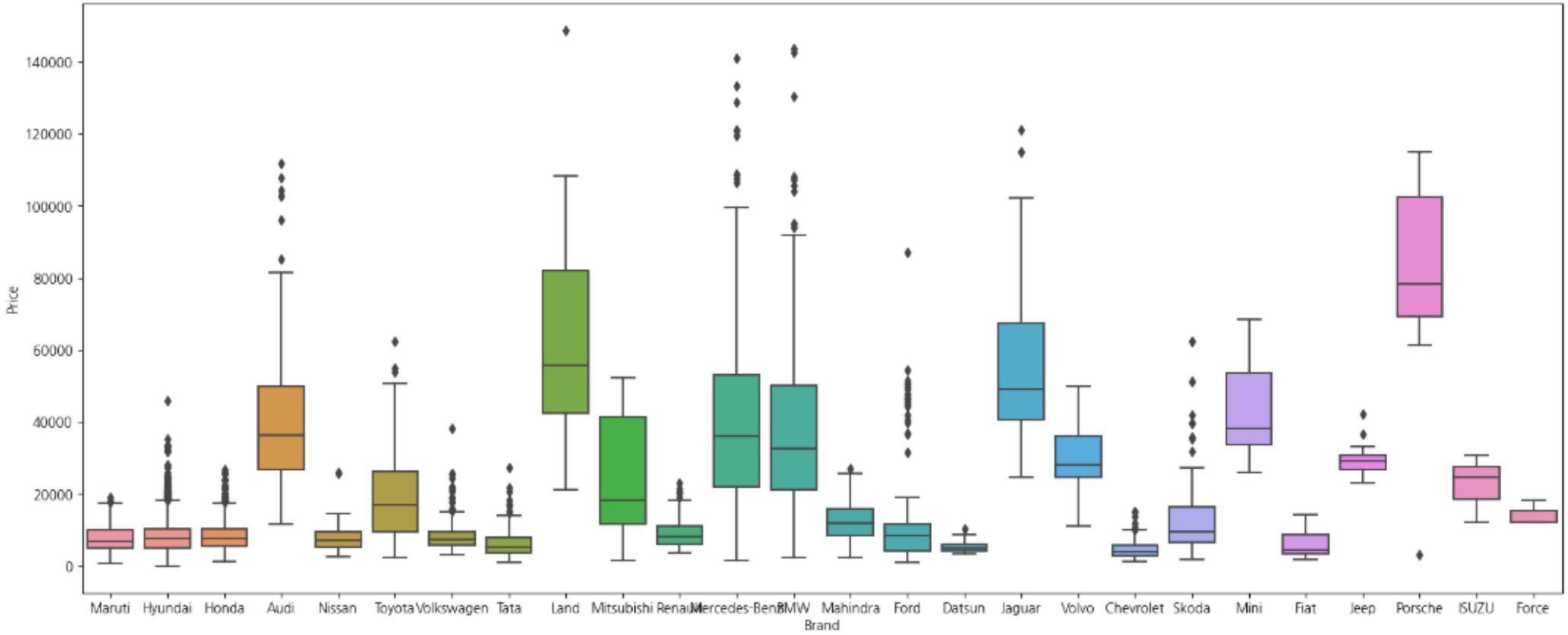
Owner_Type과 Price의 분포



직접 소유일 경우에 높은 가격이 형성되어 있고, 그 외인 경우에 가격이 더 낮다
boxplot으로만 판단하기 어려워, 검정을 수행하여 판단하기로 함

- 범주형 설명변수 항목과 목표변수 간의 Boxplot

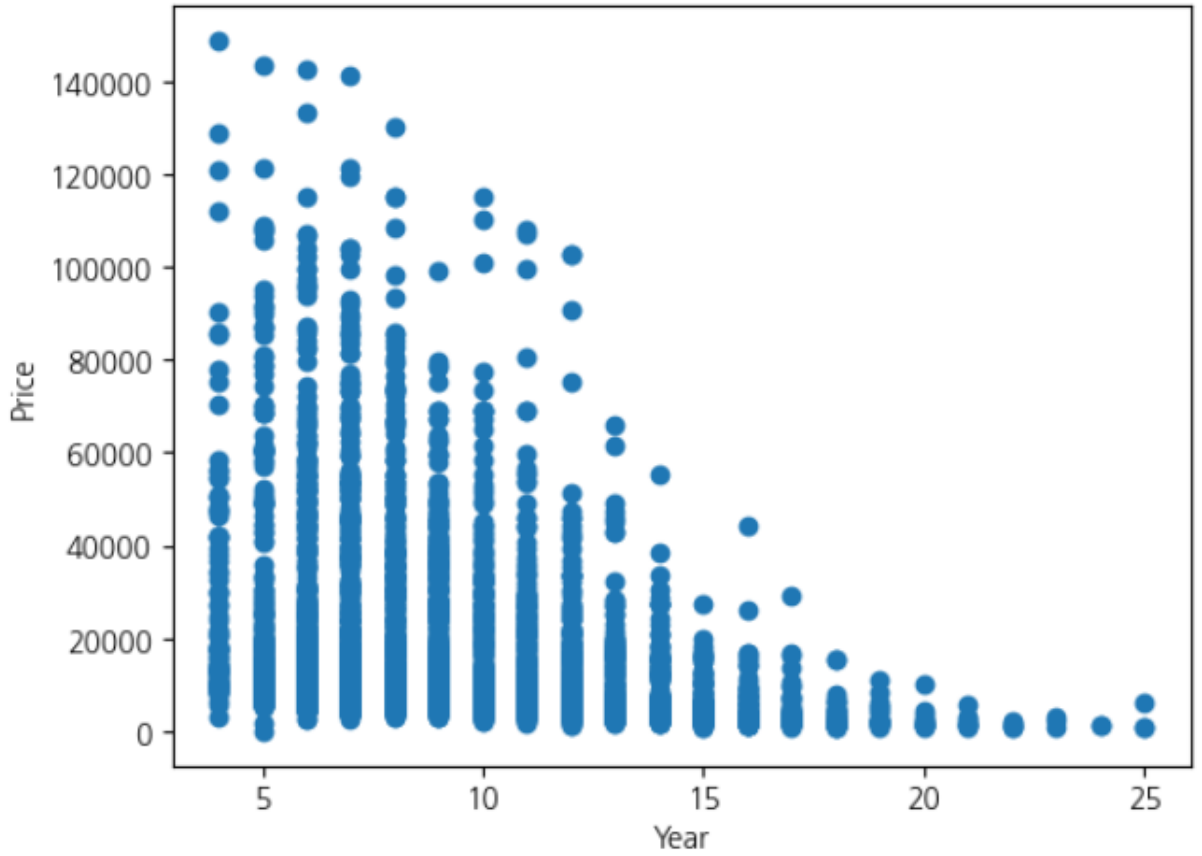
Brand과 Price의 분포



그래프를 보면 많은 관계가 있다고 할 순 없지만, 자동차의 브랜드마다 가격차이가 나는 경우가 많기 때문에 관계성이 있다고 판단하였다.

- 연속형 설명변수 항목과 목표변수 간의 Scatter plot

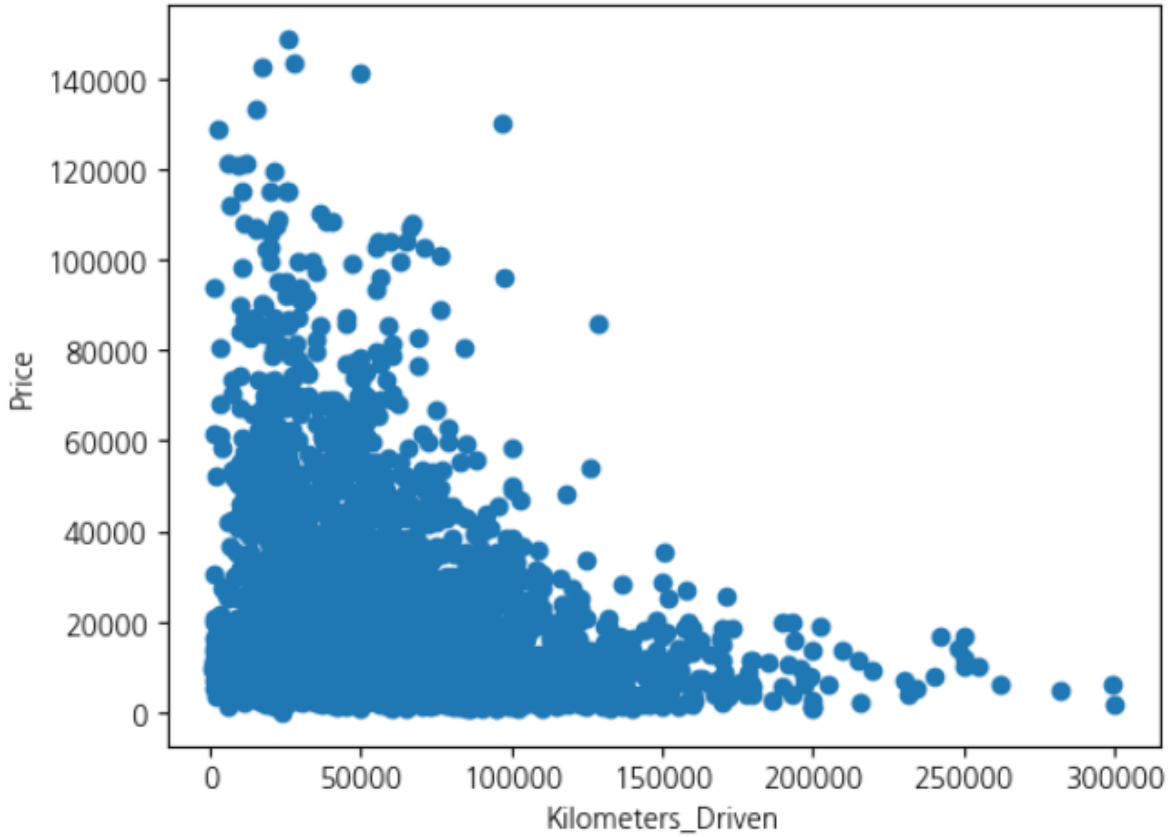
Year과 Price의 분포



연식이 오래되지 않을수록 가격이 높다는 것을 확인할 수 있음

- 연속형 설명변수 항목과 목표변수 간의 Scatter plot

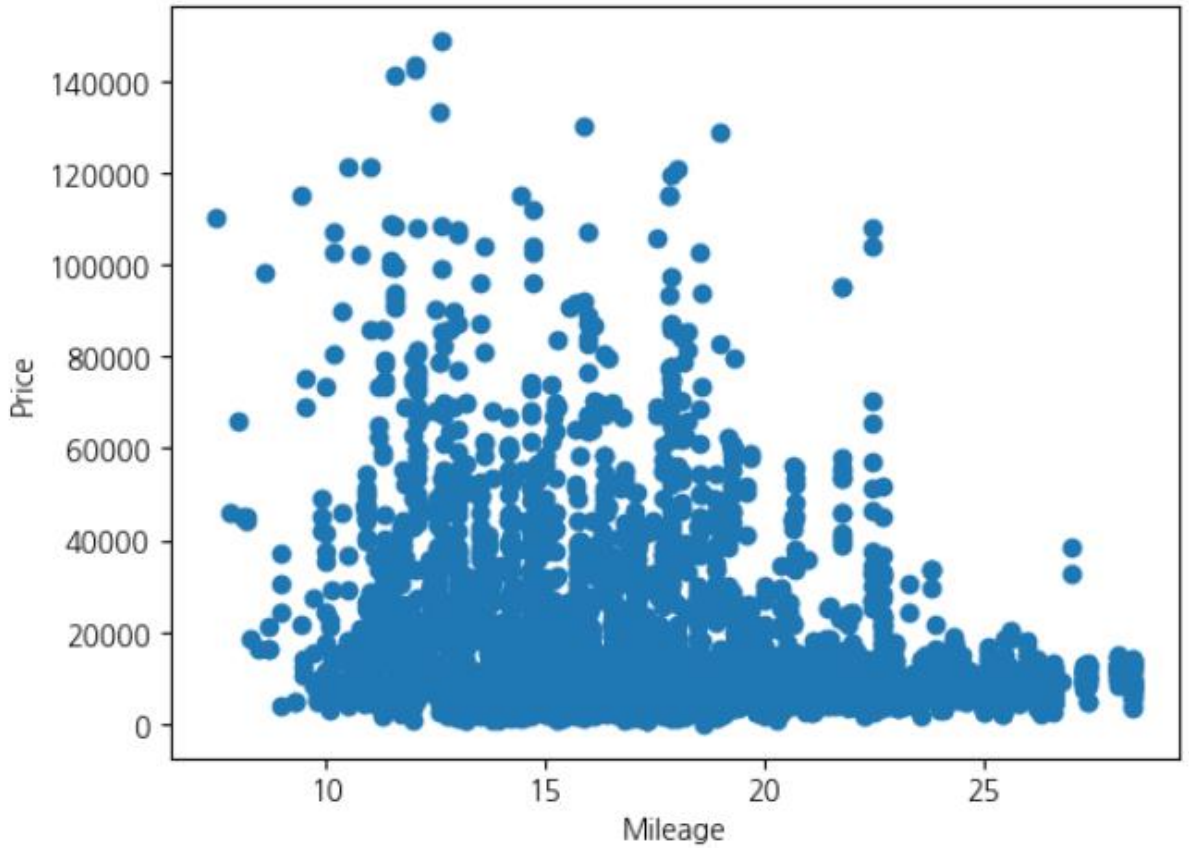
Kilometers_Driven과 Price의 분포



적은 km 수를 탈수록 가격이 높은 것을 확인할 수 있음

- 연속형 설명변수 항목과 목표변수 간의 Scatter plot

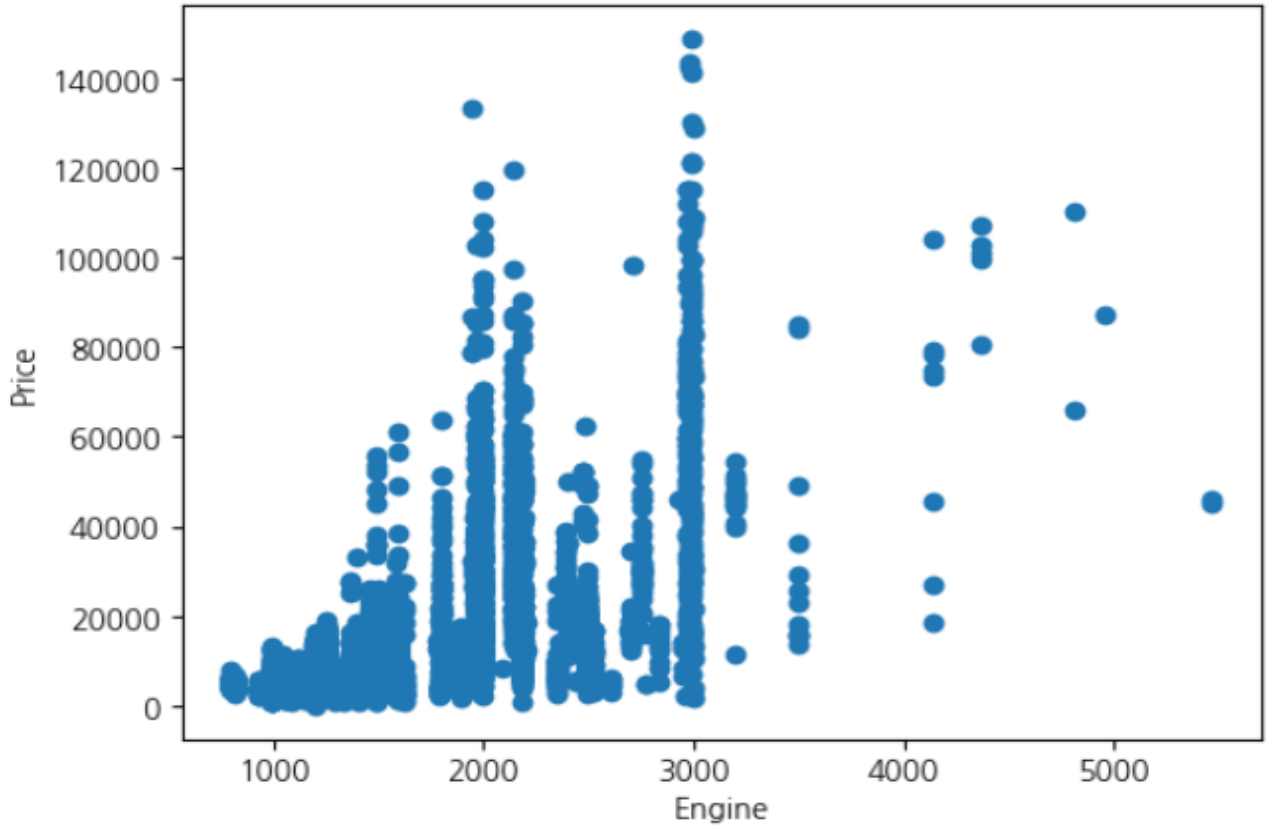
Mileage과 Price의 분포



연비에 따른 가격과의 관계는 뚜렷하지 않으므로 관계가 미미하다고 판단함.

- 연속형 설명변수 항목과 목표변수 간의 Scatter plot

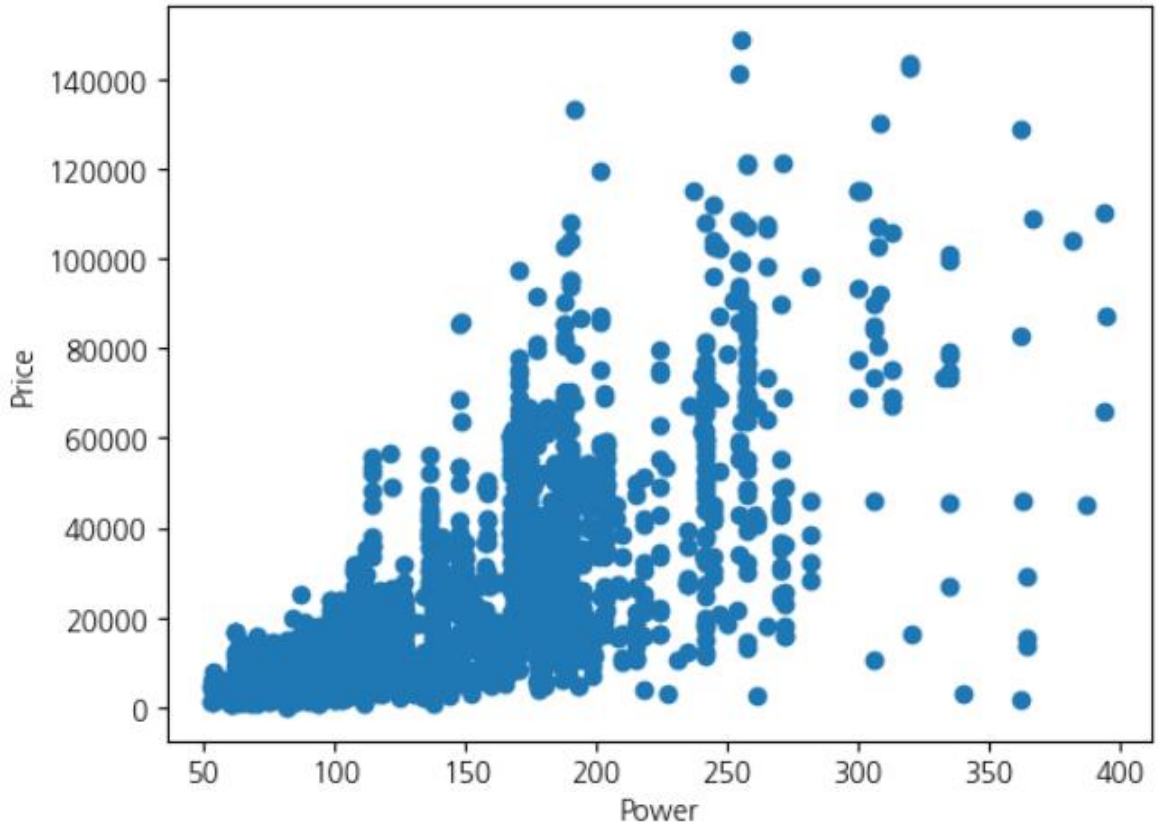
Engine과 Price의 분포



배기량이 3000cc까지는 커질수록 가격이 높아지는 경향이 있으나, 그 이후로는 뚜렷한 관계가 보이지 않음.
배기량이 클수록 엔진의 크기와 차체 크기가 큰데, 무조건적으로 큰 차가 높은 가격을 보이지는 않음

- 연속형 설명변수 항목과 목표변수 간의 Scatter plot

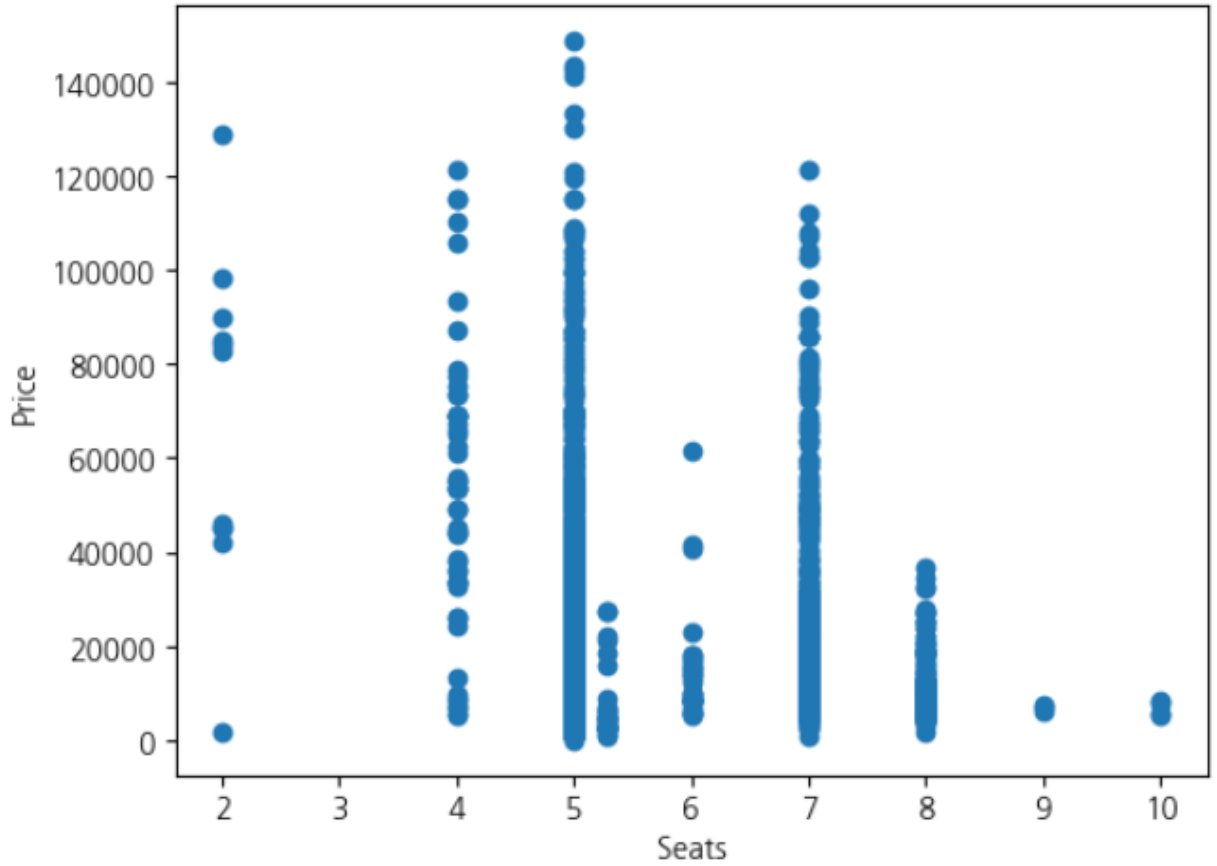
Power과 Price의 분포



최대 출력이 클수록 가격이 높다.

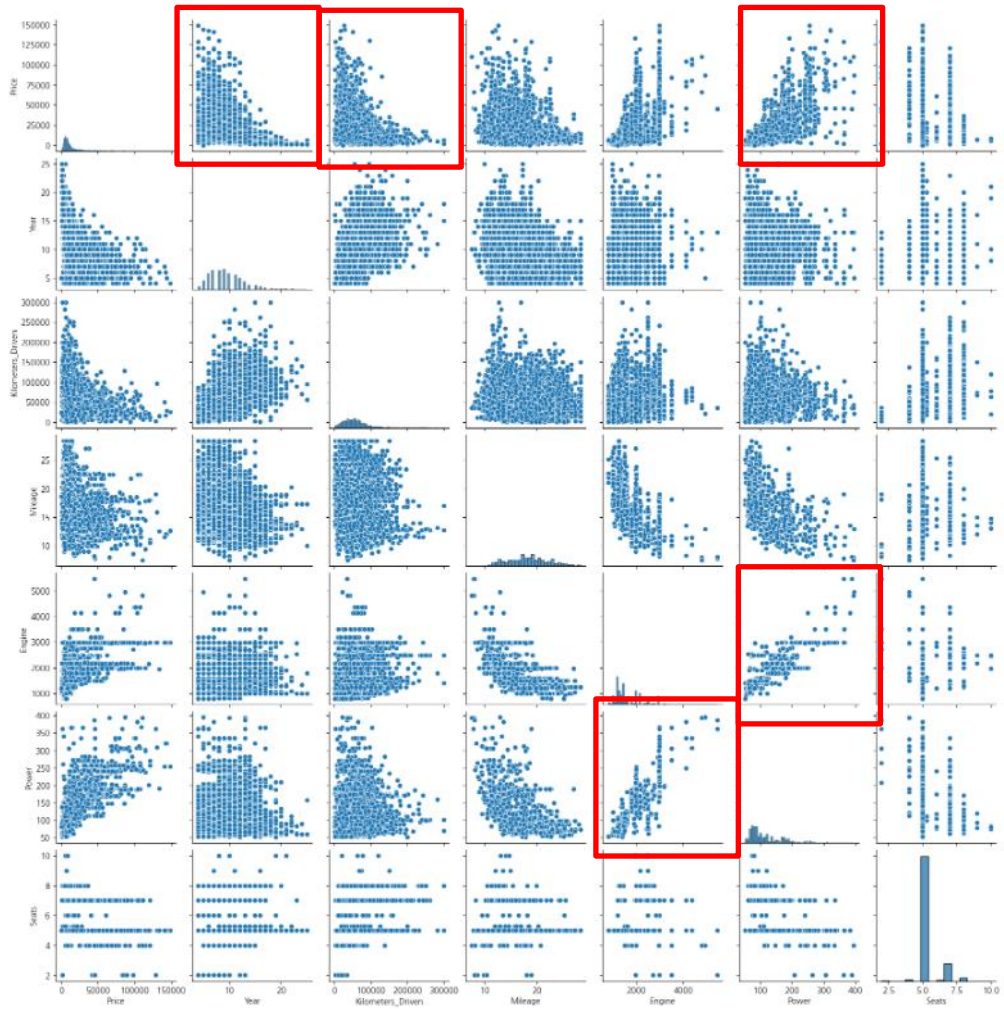
- 연속형 설명변수 항목과 목표변수 간의 Scatter plot

Seats과 Price의 분포



좌석 수와 가격은 큰 관계성이 있다고 볼 수 없다

- 연속형 설명변수 항목과 목표변수 간의 Scatter matrix



Price는 year, Kilometers_Driven, Power와 선형관계가 있을 것이다.

Power와 engine에는 서로 연관성이 있을 것이다.

- 설명변수 항목과 목표변수 간 차이 검정

자동차의 사용변속기 종류와 가격 차이를 확인하기 위한 2 smaple t-test 진행

데이터 분리

사용변속기 종류가 x = 수동일 때의 Price
사용변속기 종류가 x = 자동일 때의 Price

정규성검정

```
Shapiro-Wilk Test : statistic=0.8126437664031982, p-value=7.280306041553155e-41
Shapiro-Wilk Test : statistic=0.8679691553115845, p-value=7.065930694117904e-36
```

두 집단의 p값 < 0.05, 귀무가설이 기각되어 정규성을 띄지 않음
그럼에도 정규성을 띈다고 가정하고 등분산성 검정을 실시

등분산성검정

```
BartlettResult(statistic=2811.830724651099, pvalue=0.0)
```

t-test

```
2-sample t-test
t: -37.261
p: 0.0
```

두 집단의 p값 < 0.05, 귀무가설이 기각되어 등분산성을 띄지 않음
그럼에도 등분산성을 띈다고 가정하고 2 sample t-test검정을 실시
그렇지만 정규성, 등분산성을 둘 다 만족해야 하므로 다른 검정방식 사용

- 설명변수 항목과 목표변수 간 차이 검정

제조사별 중고차 가격 차이 확인을 위한 ANOVA 검정

```
#제조사별 중고차 가격 차이 검정
group_data = []
for group in df['Brand'].unique():
    group_data.append(df.loc[df['Brand']==group, 'Price'])

f_statistic, p_value = f_oneway(*group_data)
print("F-statistic: {}".format(f_statistic.round(3)))
print("p_value: {}".format(p_value.round()))
```

```
F-statistic: 318.972
p_value: 0.0
```

Anova 분석 결과 유의수준 5%에서 검정 결과 p 값이 0.00이므로
제조사 별 가격차이가 있다고 판단할 수 있다.

- 설명변수 항목과 목표변수 간 차이 검정

위치 별 중고차 가격 차이 확인을 위한 ANOVA 검정

```
#위치 별 중고차 가격 차이 검정
group_data = []
for group in df['Location'].unique():
    group_data.append(df.loc[df['Location']==group, 'Price'])

f_statistic, p_value = f_oneway(*group_data)
print("F-statistic: {}".format(f_statistic.round(3)))
print("p_value: {}".format(p_value.round(10)))
```

```
F-statistic: 39.349
p_value: 0.0
```

Anova 분석 결과 유의수준 5%에서 검정 결과 p 값이 0.00이므로
위치 별 중고차 가격 차이가 있다고 판단할 수 있다.

- 설명변수 항목과 목표변수 간 차이 검정

변속기 별 중고차 가격 차이 확인을 위한 ANOVA 검정

```
#변속기 별 중고차 가격 차이 검정
group_data = []
for group in df['Transmission'].unique():
    group_data.append(df.loc[df['Transmission']==group, 'Price'])

f_statistic, p_value = f_oneway(*group_data)
print("F-statistic: {}".format(f_statistic.round(3)))
print("p_value: {}".format(p_value.round()))
```

```
F-statistic: 3030.408
p_value: 0.0
```

Anova 분석 결과 유의수준 5%에서 검정 결과 p 값이 0.00이므로
변속기 별 중고차 가격 차이가 있다고 판단할 수 있다.

- 설명변수 항목과 목표변수 간 차이 검정

제조사별과 소유권 종류에 따른 카이제곱 검정

```
#제조사별과 소유권 종류에 따른 카이제곱 검정
df_b = pd.crosstab(df['Brand'],df['Owner_Type'])

chi, pval, dof, expected = stats.chi2_contingency(df_b)

print("Chisq: {}".format(chi.round(3)))
print("p_value: {}".format(pval.round()))
print("Degree of freedom: {}".format(dof))
print("expected value: {}".format(expected.round()))
```

```
Chisq: 117.398
p_value: 0.0
```

카이제곱 검정 결과 유의수준 5%에서 검정 결과 p 값이 0.00이므로
제조사별로 소유권 종류의 차이가 있다고 판단할 수 있다.
그러므로, 제조사와 소유권 사이의 영향력이 있다.

- 연속형 변수간의 상관관계 확인

산점도에서 확인한 것처럼 Power, Engine이 상관 정도가 높다.
목표변수 Price에 대한 상관관계 중, Seats를 제외한 나머지 칼럼들도 상관관계가 낮지만,
도메인 지식을 바탕으로 중요하다고 생각함

다중선형 회귀분석

연속형 변수사용

"Price", "Year" ,"Power", "Kilometers_Driven", "Mileage", "Engine"

유의성 검정

'Year', 'Power','Kilometers_Driven'

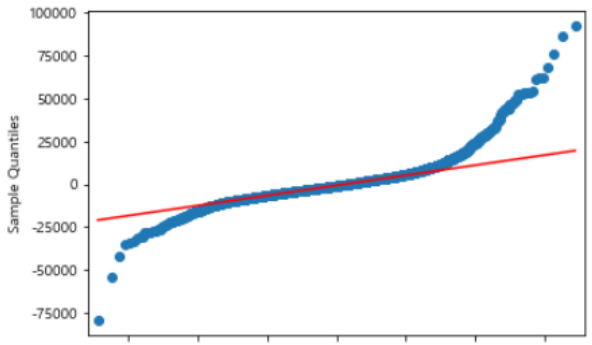
다중공선성 확인

	variable	VIF
2	Power	1.00
3	Kilometers_Driven	1.34
1	Year	1.34
0	const	15.54

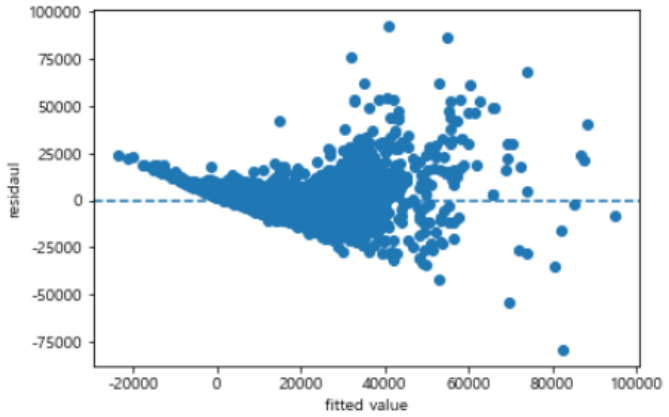
모든 변수의 다중공성선 지수 <10

- 오차의 기본가정확인

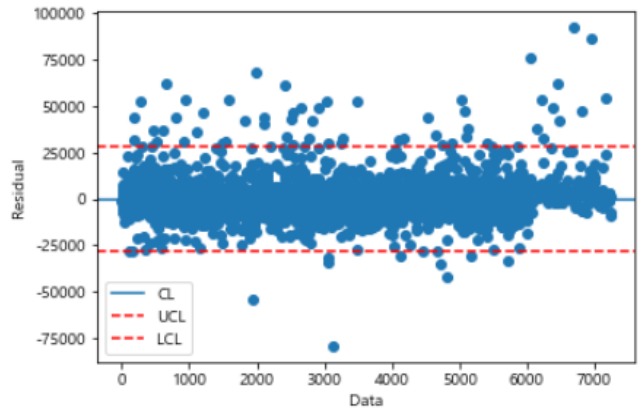
정규성 검정



등분산성 검정



독립성 검정



독립성 검정시, 잔차가 관리상하한을 많이 벗어나기 때문에, 해당 데이터를 다중선형회귀분석 모델링하기에 적합하지 않다고 판단

의사결정나무 1차 모델링

연속형 변수사용,
범주형변수 더미화 진행

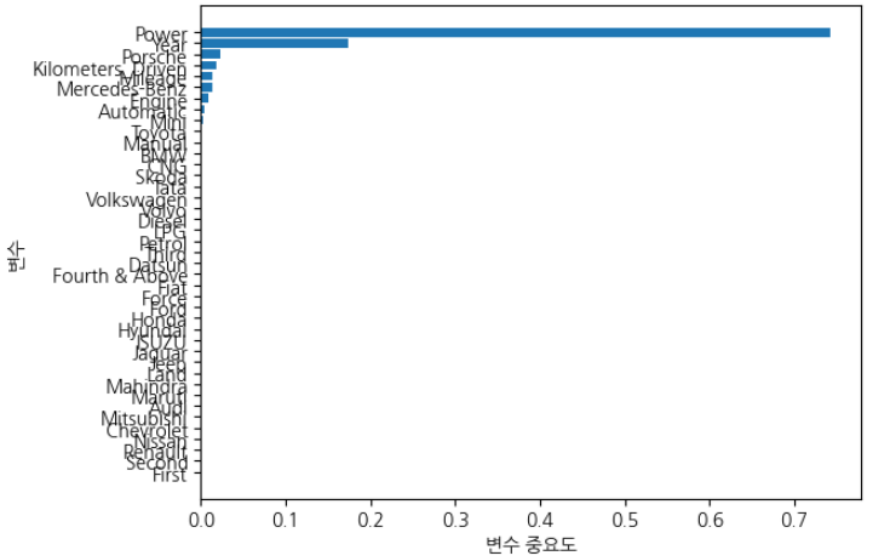
하이퍼 파라미터 튜닝

(random_state=1234,min_samples_leaf=4,
min_samples_split = 8, max_depth = 6)

1차 모델링 설명계수

```
Score on training set: 0.872649
Score on test set: 0.798608
```

변수 중요도



변수중요도는 Power.Year.Porsche.Kilometers_Driven, Mileage
순으로 목표변수인 중고차 가격 Price에 영향이 크다고 해석

비교적 낮은 중요도인 Fuel_Type컬럼을 제거하고 다시 2차 모델링 진행이 필요.

의사결정나무 2차 모델링

연속형 변수사용,
범주형변수 더미화 진행

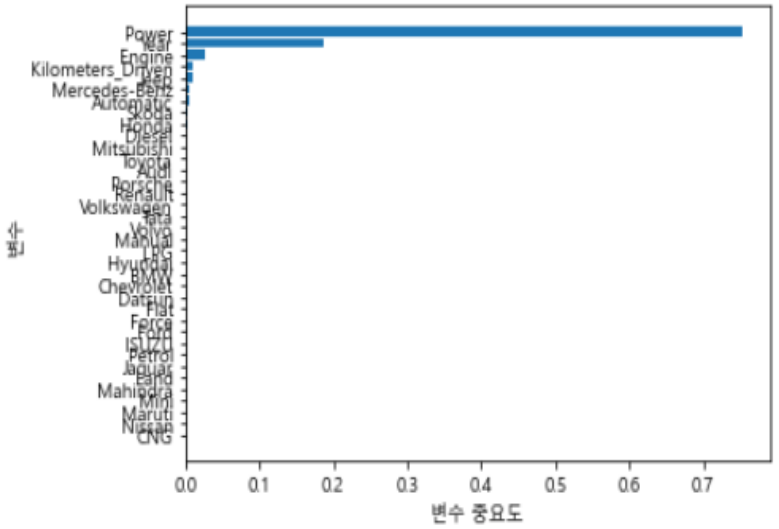
하이퍼 파라미터 튜닝

(random_state=1234,min_samples_leaf=4,
min_samples_split = 8, max_depth = 6)

2차 모델링 설명계수

```
Score on training set: 0.868011
Score on test set: 0.808698
```

변수 중요도



컬럼을 제거한 결과
train : 87.3% -> 86.8%
Test : 79.9% -> 80.9%

2차 모델링 결과, Train의 성능을 떨어졌지만
test의 경우 성능이 올라감.

랜덤포레스트 1차 모델링

연속형 변수사용,
범주형변수 더미화 진행

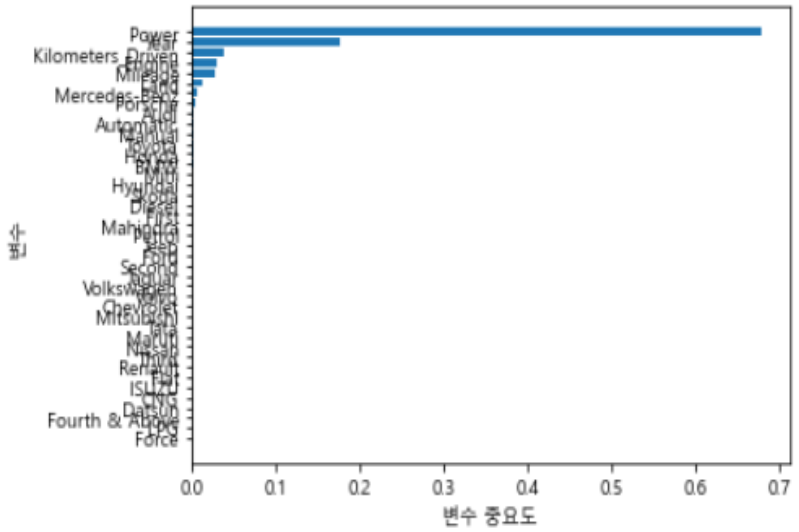
하이퍼 파라미터 튜닝

(n_estimators= 80, random_state=1234 ,
min_samples_leaf=1,min_samples_split= 4,max_depth = 10)

1차 모델링 설명계수

```
Score on training set: 0.958607
Score on test set: 0.850022
```

변수 중요도



변수중요도는 Power.Year.Kilometers_Driven, Engine, Mileage 순으로
목표변수인 중고차 가격 Price에 영향이 크다고 해석

비교적 낮은 중요도인 Owner_Type 컬럼을 제거하고 다시 2차 모델링 진행이 필요.

랜덤포레스트 2차 모델링

연속형 변수사용,
범주형변수 더미화 진행

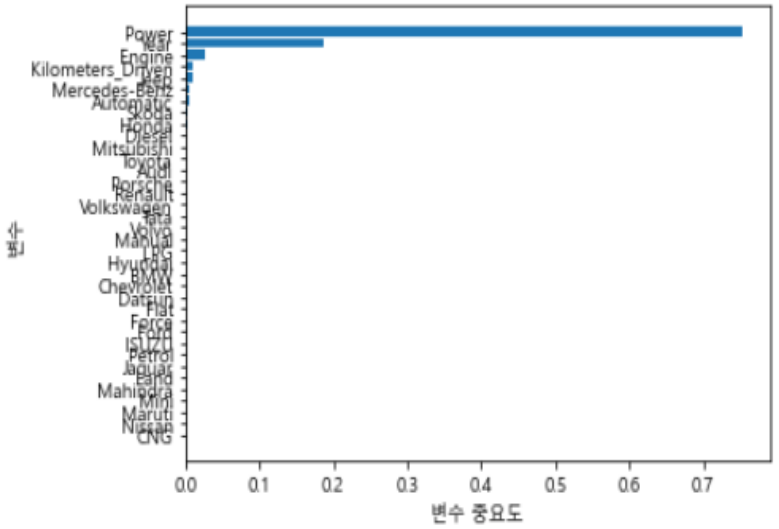
하이퍼 파라미터 튜닝

(random_state=1234,min_samples_leaf=4,
min_samples_split = 8, max_depth = 6)

2차 모델링 설명계수

```
Score on training set: 0.958587
Score on test set: 0.850931
```

변수 중요도



컬럼을 제거한 결과
train : 95.9% -> 95.9%
Test : 85.0% -> 85.1%

2차 모델링 결과, test의 경우 성능이 올라감.

그래디언트 부스팅 1차 모델링

연속형 변수사용,
범주형변수 더미화 진행

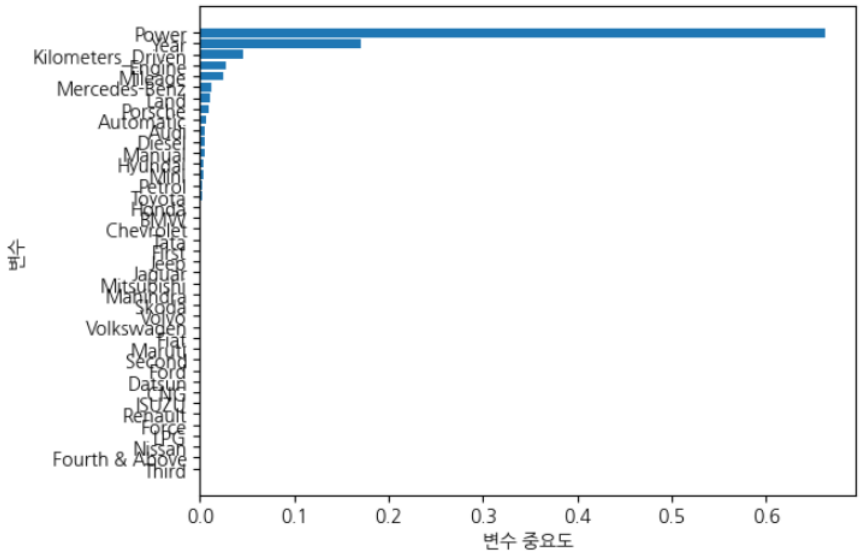
하이퍼 파라미터 튜닝

(random_state=1234, n_estimators = 140,min_samples_leaf=3,
min_samples_split=6, max_depth=4, learning_rate=0.1)

1차 모델링 설명계수

```
Score on training set: 0.960
Score on test set: 0.830
```

변수 중요도



변수중요도는 Power.Year.Kilometers_Driven, Engine, Mileage 순으로
목표변수인 중고차 가격 Price에 영향이 크다고 해석

비교적 낮은 중요도인 Owner_Type 컬럼을 제거하고 다시 2차 모델링 진행이 필요.

그래디언트 부스팅 2차 모델링

연속형 변수사용,
범주형변수 더미화 진행

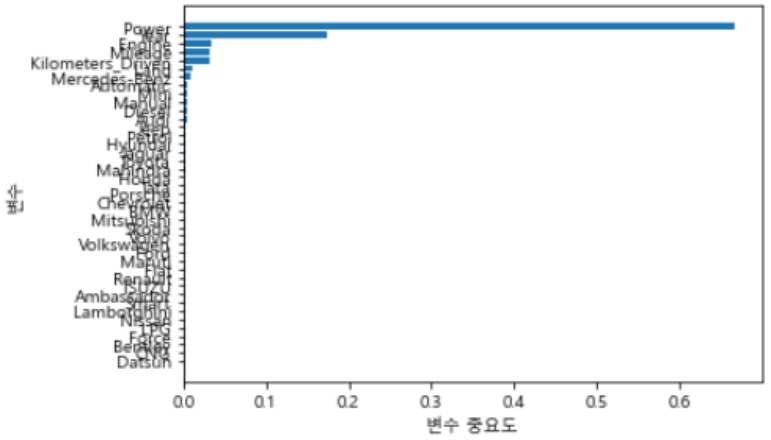
하이퍼 파라미터 튜닝

(random_state=1234, n_estimators = 140,min_samples_leaf=3,
min_samples_split=6, max_depth=4, learning_rate=0.1)

2차 모델링 설명계수

```
Score on training set: 0.973
Score on test set: 0.831
```

변수 중요도



컬럼을 제거한 결과
train : 96.0% -> 97.3%
Test : 83.0% -> 83.1%

2차 모델링 결과, train, test의 경우 성능이 올라감.

다중공선성 확인

	variable	VIF
0	const	0.00
2	Kilometers_Driven	1.04
1	Year	1.37
3	Mileage	2.86
5	Power	6.91
4	Engine	8.83
26	Mitsubishi	inf
27	Nissan	inf
28	Porsche	inf
29	Renault	inf
30	Skoda	inf
31	Smart	inf
32	Tata	inf
35	Volvo	inf
34	Volkswagen	inf
25	Mini	inf
36	CNG	inf
37	Diesel	inf
38	LPG	inf
39	Petrol	inf
33	Toyota	inf
24	Mercedes-Benz	inf
21	Land	inf
22	Mahindra	inf
6	Ambassador	inf
7	Audi	inf
8	BMW	inf

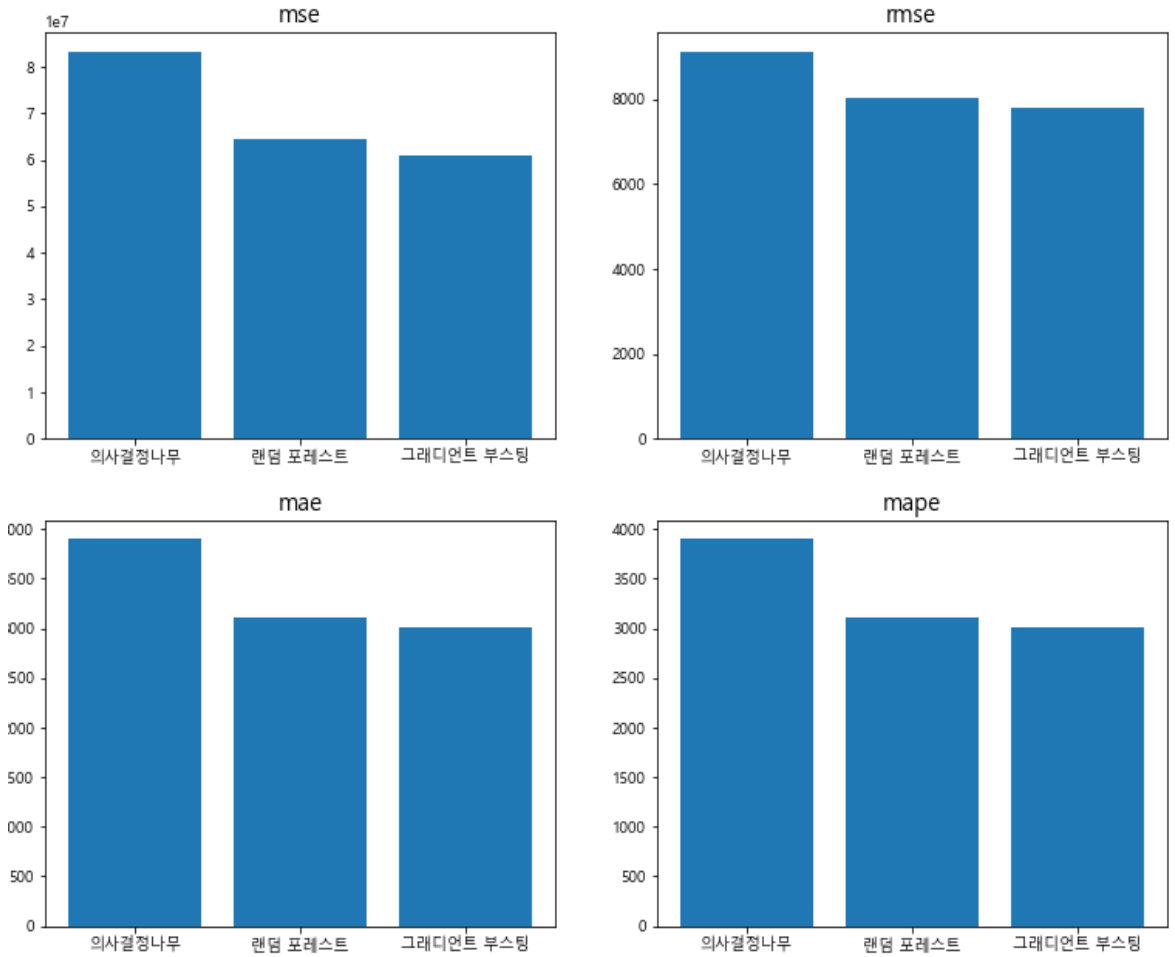
다중공선성확인 결과, 연속형 변수인 경우 VIF지수가 10이하로 괜찮음.

그러나 범주형 변수인 경우 더미화를 했기 때문에 inf로 표시됨.

이는 더미화를 함으로써 자연스럽게 생기는 문제라고 판단을 함.

그러므로 전부 의미있는 컬럼이라고 판단하여 제외시키지 않고 컬럼을 그대로 사용

모델별 성능지표 확인



성능지표 확인 결과, 그래디언트 부스팅 > 랜덤 포레스트 > 의사결정나무 순으로 좋게 측정 되었다.

핵심인자 선정 및 결과 해석

회귀분석, 의사결정나무, 랜덤 포레스트, 그래디언트 부스팅 모델링 결과
목표변수인 Price에 많은 영향을 미치는 인자는 Year, Power, Kilometer_Driven이다
또한 범주형 설명변수 중에서도, Brand, Transmission이 영향을 많이 미쳤다.

영향을 많이 미치는 인자를 정리해보면
Power > Year > Brand > Killometer_Driven 순으로 정리할 수 있다.

핵심인자 정리 템플릿(참조용)

핵심인자 선정을 위한 분석 과정에서 나온 결과를 순위 등으로 종합 정리합니다.
각자 필요한 형식으로 변경해서 사용하세요(엑셀 파일 제공)

변 수	변수 설명	변수 역할	변수 형태	분석제외 사유	탐색적 기법			모델링 기법						총점	선정 (사유)
					그래프	t/F 검정	X2 검정	회귀 분석	DT	RF	GB	…	기술적 검토		
Price	중고차 가격(단위:천원)	목표변수	연속형		o			o	o	o	o			목표	
Name	자동차 브랜드와 모델	설명변수	연속형	브랜드로 변환		o	o		o	o	o			0.2	선호브랜드 존재
Location	자동차를 팔거나 구매할 수 있는 위치	설명변수	범주형	boxplot결과 비슷함.		o									
Year	모델의 년도 혹은 버전	설명변수	연속형		o			o	o	o	o			17.8	변수중요가 높음
Kilometers_Driven	이전 소유주의 차량 주행거리(Km)	설명변수	연속형		o			o	o	o	o			3.8	연식과 관련
Fuel_Type	자동차 사용연료	설명변수	범주형		o				o	o	o			0.01	선호도가 존재
Transmission	자동차 변속기 종류	설명변수	범주형			o			o	o	o			0.01	선호도가 존재
Owner_Type	소유권이 직접 소유인지, 중고 소유인지 여부	설명변수	범주형	큰차이가 없음			o								
Mileage	자동차 회사가 제공하는 표준 주행거리(kmpl)	설명변수	연속형		o			o	o	o	o			2.8	연식과 관련
Engine	엔진의 배기량(cc)	설명변수	연속형		o			o	o	o	o			2.9	차 성능과 관련
Power	엔진의 최대 출력(bhp)	설명변수	연속형		o			o	o	o	o			67.8	변수중요도가 높음
Seats	차의 좌석 수	설명변수	연속형	큰차이가 없음	o										
New_Price	뉴모델의 가격	설명변수	연속형	결측치가 많음											