중고차 가격 예측 모델 개발 및 모델 성능 향상 방안

A2 김승희

과제 정의

인도 중고차 시장의 지속적인 성장세

핵심영향인자 도출

중고차 가격에 영향을 주는 변수 찾기

가격예측모델

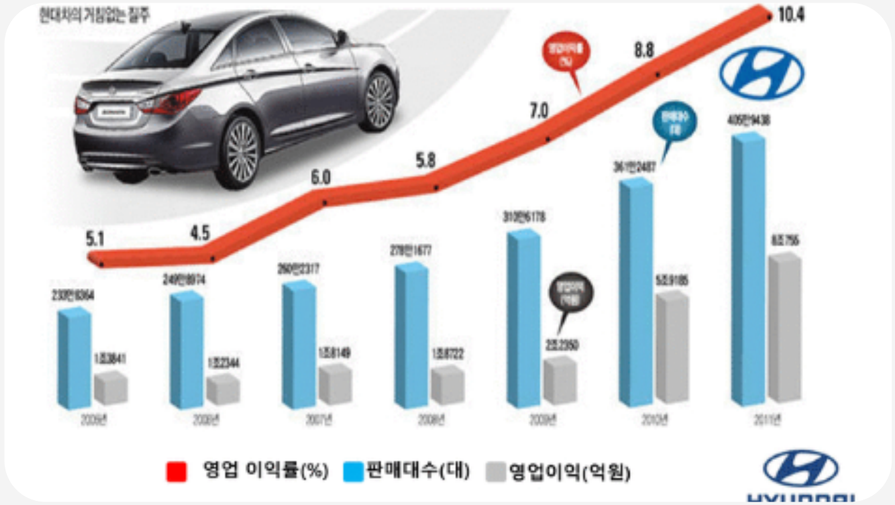
새로운 중고차 데이터가 들어왔을 때
가격을 잘 예측할 수 있는 모델 찾기

인도 중고자 시장에서 경쟁력 확보, 수익성 향상

사전 조사

인도 자동차 시장 현황

일본	인도	독일
484만대(3)	370만대(4)	347만대(5)
497만대(3)	367만대(5)	382만대(4)



국가별 자동차 산업 수요

2023년 기준 인도는 작년 자동차 판매에서 일본을 추월하여 세 번째로 큰 자동차 시장

인도에서 국내브랜드 자동차 판매량 추이

억눌린 수요와 공급망 부족의 완화로 연료 및 자동차 가격 인플레이션, 금리 강세, 금융위기 등 글로벌 역풍에도 불구하고 2022년에는 역대 최고 상용차 판매량을 기록

https://www.motorgraph.com/news/articleView.html?idxno=14096
https://www.kita.net/cmmrcInfo/cmmrcNews/overseasMrktNews/overseasMrktNewsDetail.do;JSESSIONID_KITA=37B10C8DD905C68742C49E012CBAE25B.Hyper?pageIndex=1&nIndex=1830795&type=0

분석 계획

데이터 확인

- 결측치, 및 이상치 제거

그래프를 통한 탐색적 분석

- 범주형 설명변수에서 가격 차이가 있는지
- 설명변수 간 상관성 분석

모델 생성

- 회귀분석, DecisionTree, RandomForest, Gradient Boosting 모델 생성 및
- 모델별 지표 비교
- 모델 평가 및 평가지표 선택

데이터 분석 전 예상

1. 브랜드 별로 중고차 가격의 평균이 다를 것이다.
2. 위치가 도심에 가까울 수록 중고차 가격이 높을 것이다.
3. 모델의 년도가 최신일수록 중고차 가격이 높을 것이다.
4. 보통 연식대비 주행거리가 긴 중고차는 좋은 중고차라는 인식이 있다 (연식은 최신이지만 주행거리가 긴 차량은 고속도로를 주로 주행한 차량이고 고속도로는 시내 주행에 비해 브레이크를 밟는 횟수가 현저히 적기 때문에 과부하가 적다) 따라서 연식대비 주행거리가 긴 중고차일수록 중고차의 가격이 높을 것이다.
5. 인도 정부 BS-VI 배출가스 규제에 따라 디젤 차량 판매 시 가솔린 차량보다 2배 높은 환경부담금을 내야 한다. 특히 올해 예정된 2단계 규제 적용 시 비용 부담은 더욱 높아진다. 따라서 디젤차량이 다른 차량보다 가격이 낮을 것이다.
6. 배기량의 크기는 엔진의 직접적인 힘인 토크와 최대출력을 나타내는 마력이 그만큼 비례하여 높아진다고 사전정보에 따라 배기량과 마력 사이에 상관관계가 있을 것으로 예상된다.
7. 자동차 사용 변속기의 경우 점점 automatic으로 출시되는 차량이 증가하는 추세이므로 manual보다 automatic의 가격이 높을 것이다.
8. 차의 좌석 수가 클수록 차량이 크다는 의미이므로 좌석 수가 많을 수록 중고차 가격이 높을 것이다.

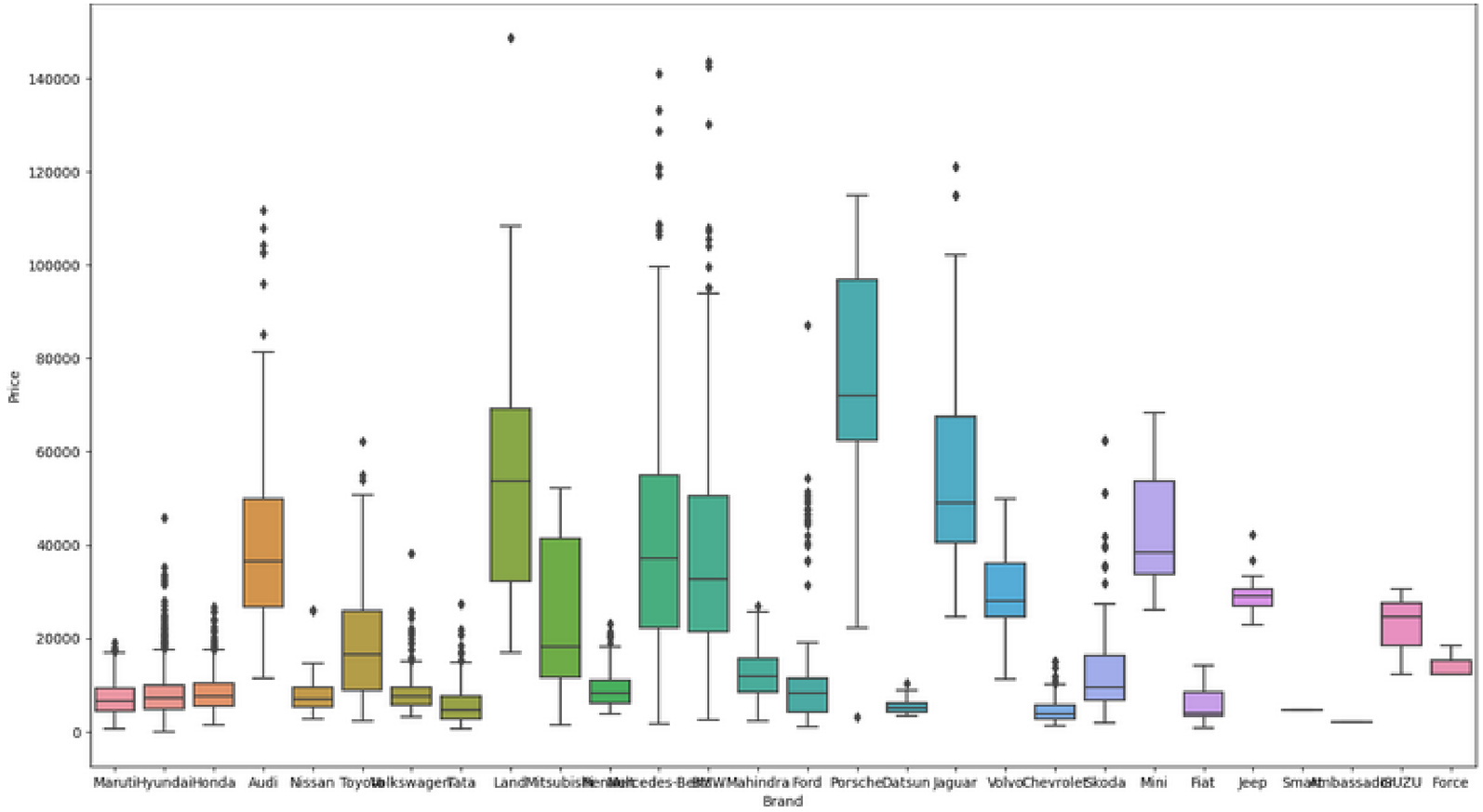
데이터 확인 및 처리

- Name은 브랜드와 모델의 이름이 함께 있으므로 Brand만 추출
- 자동차 모델의 경우 너무 다양하므로 분석의 의미가 없을 것으로 추정되어 열 삭제
- Mileage, Engine, Power는 단위와 분리하여 수치형 변수로변환
- df_raw.info()로 확인 결과 New_Price는 전체 행 중 값이 있는 행이 1006개 밖에 없음을 확인,
  데이터 수가 너무 작아서 제거
- 목표변수인 Price 1053개의 결측치 발견됨, 결측치인 행 제거
- 수치형 설명변수 Mileage, Engine, Power, Seat는 각각 2,46, 46, 53개의 결측치 발견 ,
  개수가 적으므로 평균값으로 대체
- boxplot을 그려 변수의 대략적인 분포를 확인한 뒤 이상치 제거

탐색적 분석

범주형 설명변수에 따른가격 차이

1. 브랜드 별로 중고차 가격의 평균이 다를 것이다.

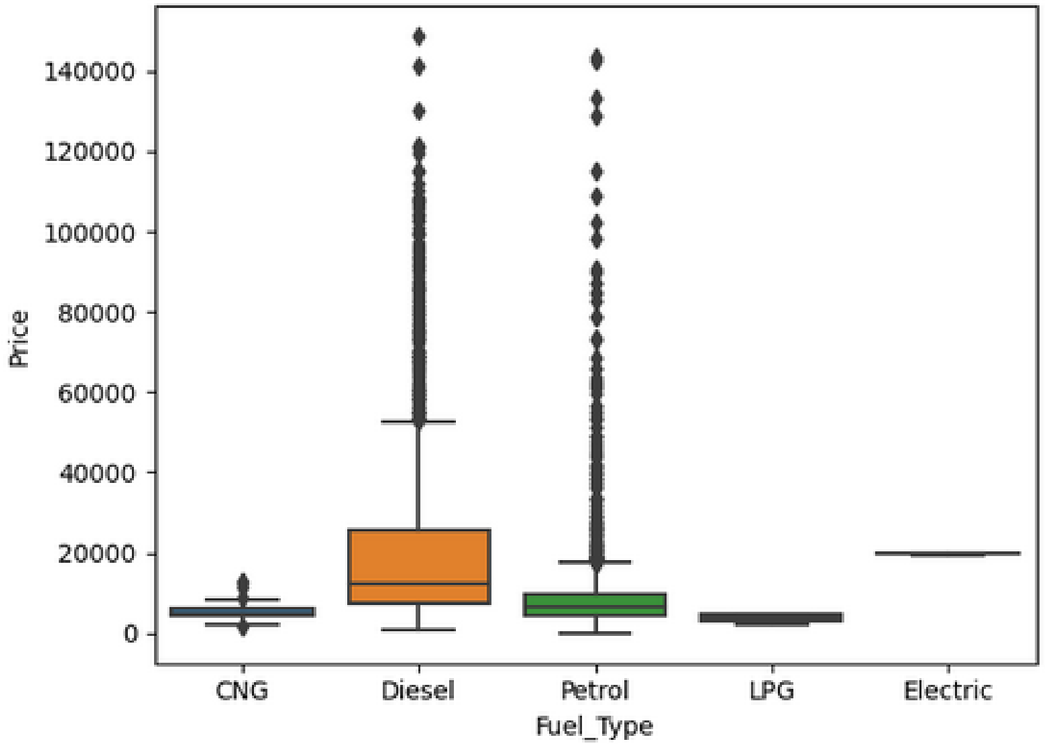


box plot을 그려서 브랜드별 로 중고차 가격의 평균을 확인해본 결과 브랜드 별로 평균이 크게 차이나는 것을 확인할 수 있다.

탐색적 분석

범주형 설명변수에 따른가격 차이

2. 연료 종류에 따른 가격 차이가 있을 것이다.




```
# ANOVA 분석 수행하기
f_statistic, p_value = stats.f_oneway(fuel_1, fuel_2,fuel_3,fuel_4,fuel_5)

print("F-statistic: {:.3f}".format(f_statistic))
print("p-value: {:.15f}".format(p_value))
```

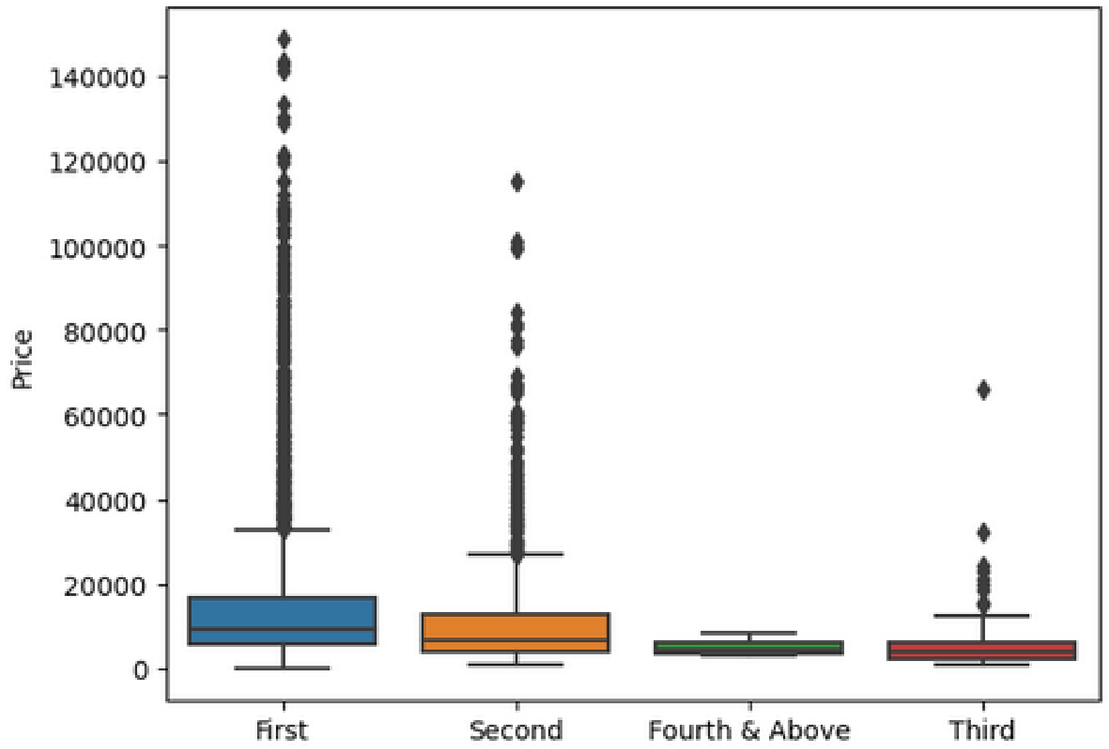
```
F-statistic: 191.767
p-value: 0.000000000000000
```

box plot을 그려서 연료 종류에 따른 중고차 가격의 평균을 확인해 보았으나 그래프만 보았을 때,
실제로 차이가 있는지 없는지 알기 어려워 ANOVA 분석 수행
=> p-value가 0.05보다 작음 => 연료 종류에 따른 중고자 가격 평균에는 차이가 있다

탐색적 분석

범주형 설명변수에 따른가격 차이

3. 소유권에 따른 가격차이가 있을 것이다.




```
# ANOVA 분석 수행하기
f_statistic, p_value = stats.f_oneway(owner_1, owner_2, owner_3, owner_4)

print("F-statistic: {:.3f}".format(f_statistic))
print("p-value: {:.15f}".format(p_value))
```

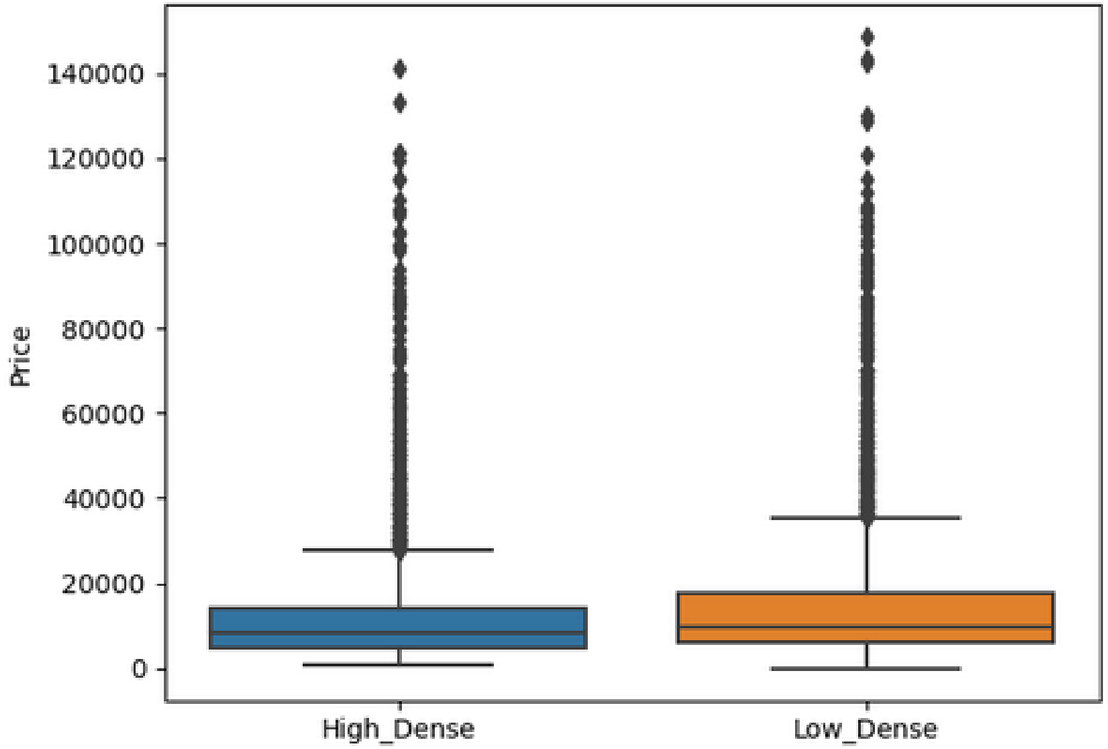
```
F-statistic: 25.663
p-value: 0.000000000000000
```

box plot을 그려서 소유권에 따른 중고차 가격의 평균을 확인해 보았으나 그래프만 보았을 때,
실제로 차이가 있는지 없는지 알기 어려워 ANOVA 분석 수행
=> p-value가 0.05보다 작음 => 소유권에 따른 중고자 가격 평균에는 차이가 있다

탐색적 분석

범주형 설명변수에 따른가격 차이

4. 지역(인구 밀도가 높은 지역 vs 인구 밀도가 낮은 지역) 간 가격 차이가 있을 것이다




```python
# 2 sample t-test
t_statistic, p_value = stats.ttest_ind(loc_0, loc_1)

print("T-statistic: {:.3f}".format(t_statistic))
print("p-value: {:.15f}".format(p_value))

df_raw = df_raw.drop('loc', axis = 1)
```

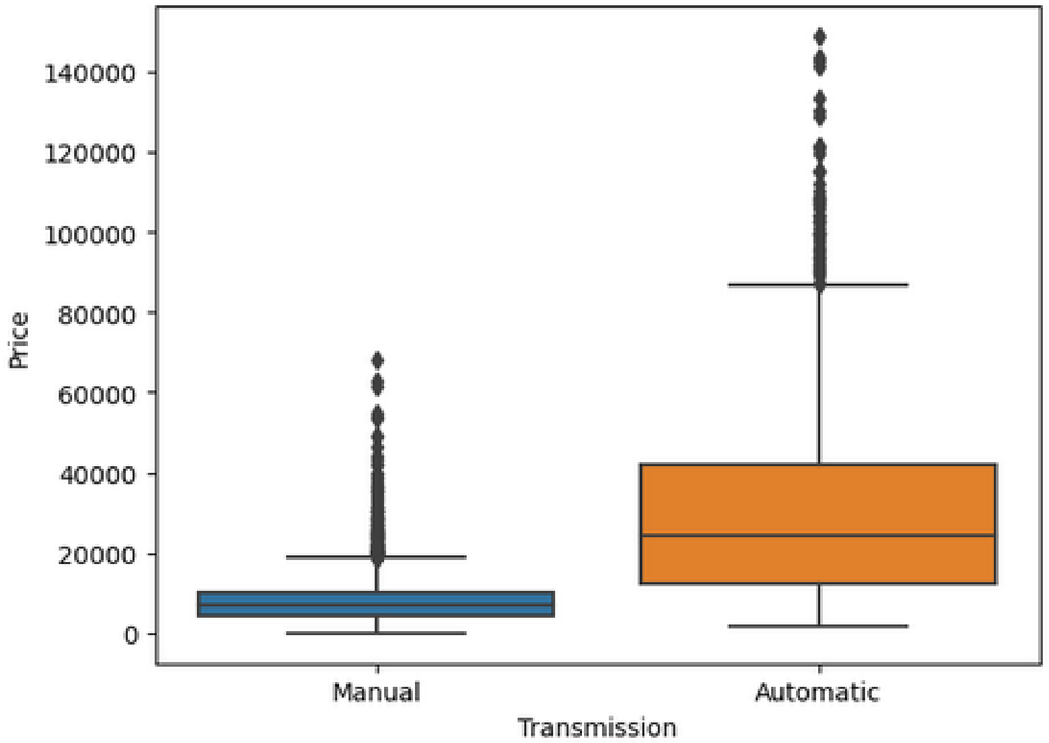
```
T-statistic: -6.345
p-value: 0.000000000237937
```

high_density_regions은 2023년 기준 높은 인구 밀도를 가진 지역 상위 5개이다.
box plot을 그려서 인구 밀도에 따른 중고차 가격의 평균을 확인해 보았으나 그래프만 보았을 때,
실제로 차이가 있는지 없는지 알기 어려워 t_test 시행 => p-value가 0.05보다 작음
=> 지역에 따른 중고자 가격 평균에는 차이가 있다

탐색적 분석

범주형 설명변수에 따른가격 차이

5. Transmission 별 가격 차이가 있을 것이다.



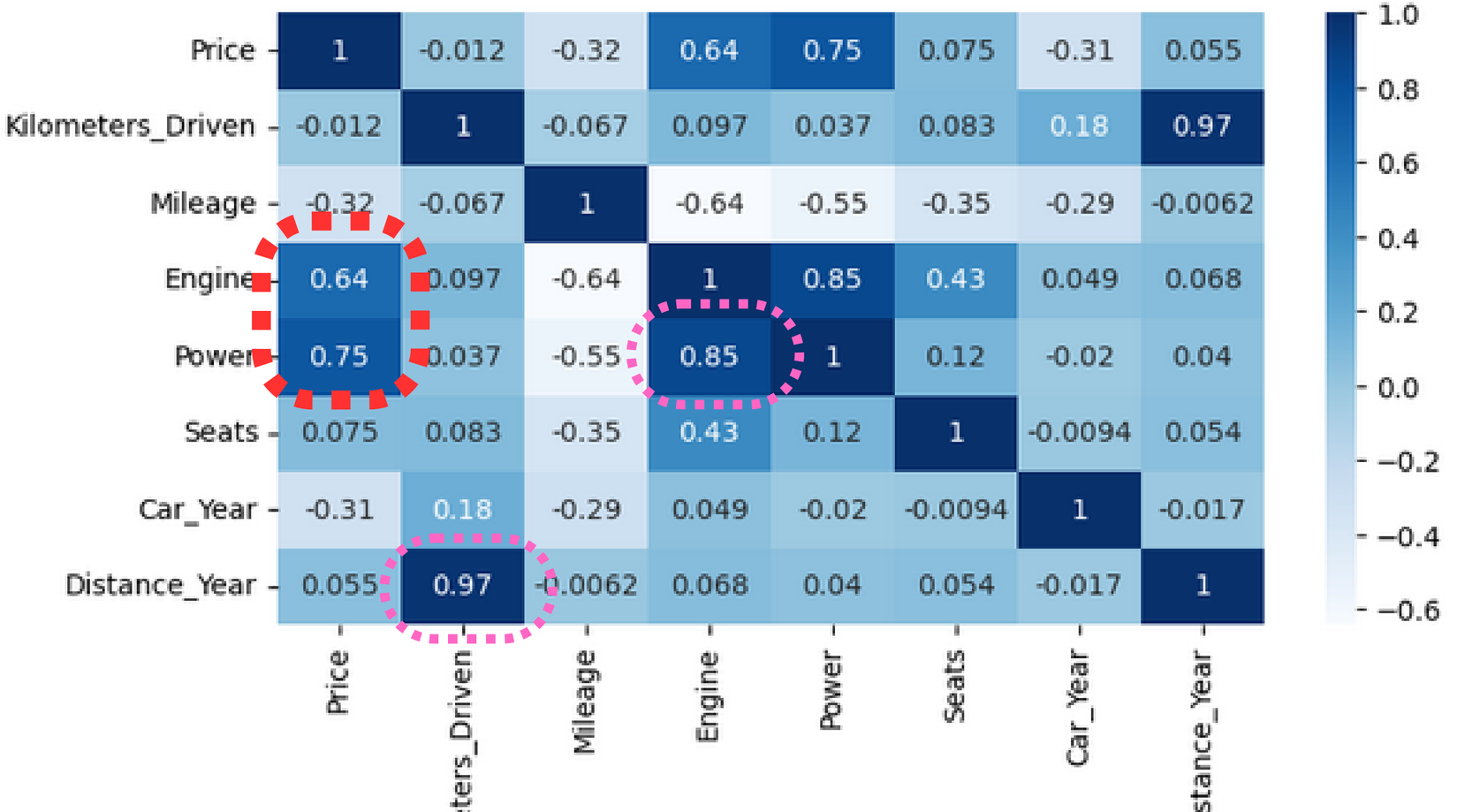

```
# 2 sample t-test
t_statistic, p_value = stats.ttest_ind(trans1, trans2)

print("T-statistic: {:.3f}".format(t_statistic))
print("p-value: {:.15f}".format(p_value))
```

```
T-statistic: 57.504
p-value: 0.000000000000000
```

box plot을 그려서 Transmission에 따른 중고차 가격의 평균을 확인해 보았으나 그래프만 보았을 때, 실제로 차이가 있는지 없는지 알기 어려워 t_test 시행=> p-value가 0.05보다 작음
=> Transmission에 따른 중고자 가격 평균에는 차이가 있다

탐색적 분석

연속형 설명변수에 따른 가격 차이



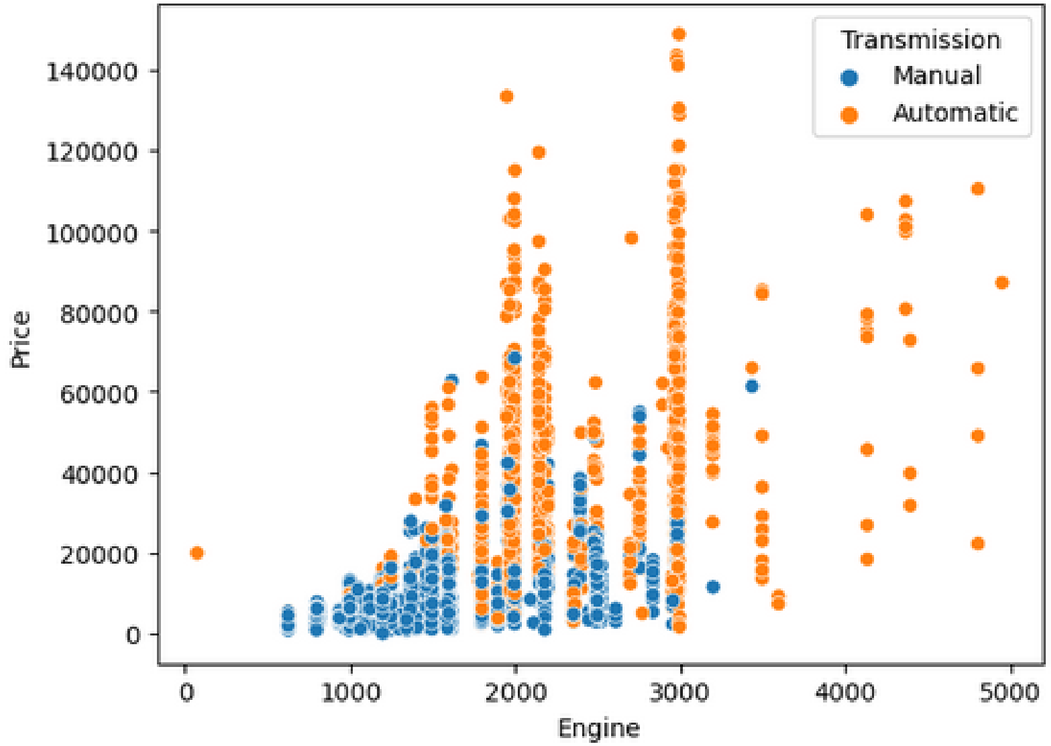
연속형 설명변수 Price는 Engine, Power와 비교적 높은 상관관계를 보인다. => 산점도를 그려 확인

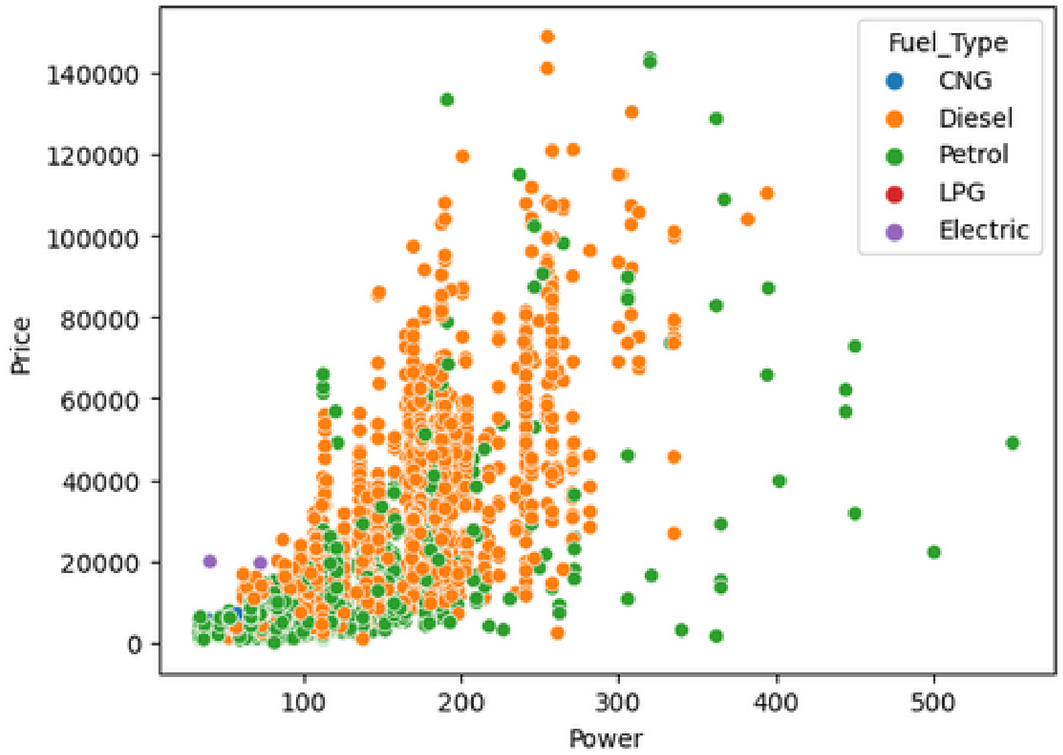
Engine과 Power 설명변수들 간 상관관계가 높다. => 추후 다중공선성 확인 필요

Distance_Year와 Kilometers_Driven 설명변수들 간 상관관계가 높다 => 추후 다중공선성 확인 필요

탐색적 분석

연속형 설명변수에 따른가격 차이





Price는 Engine, Power와 비교적 높은 상관관계를 보였기 때문에 산점도를 그려 그래프의 분포를 확인한다.
Engine과 Power 모두 값이 증가 함에 따라 Price값도 증가하는 양상을 보인다.
Engine-Price 그래프의 경우 Engine이 작을수록 Transmission에는 Manual이 많고
Power-Price 그래프의 경우 Power가 작을수록 Fuel_Type에는 Petrol이 많다.

모델 생성

회귀분석

Price ~ Brand

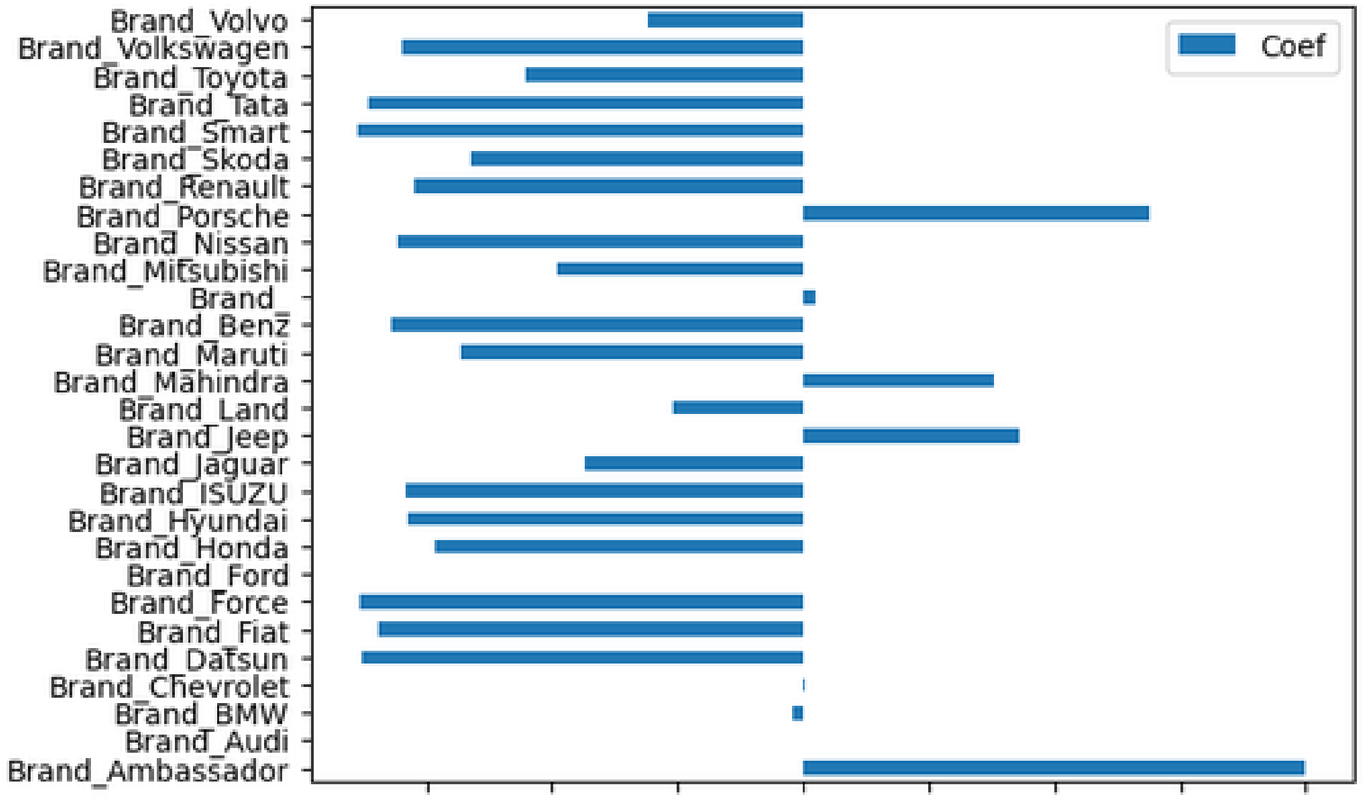
get_dummies를 사용하여 Brand변수를 분리하였다.

다중선형회귀분석 모델의 설명변수로 Brand만 넣고 실행한 결과 Brand별로 설명변수 중요도가 크게 달랐다.

이전 ANOVA를 통해 Brand별로 가격에 주는 영향이 다르다는 결과와 부합하는 모습이다.

Porche, Jeep, Jaguar, Ambassador등의 경우 Price에 양의 방향으로 영향을 준다.

반면 Hyundai, Toyota, Nissam등 의 경우 Price에 음의 방향으로 영향을 준다.



모델 생성

회귀분석

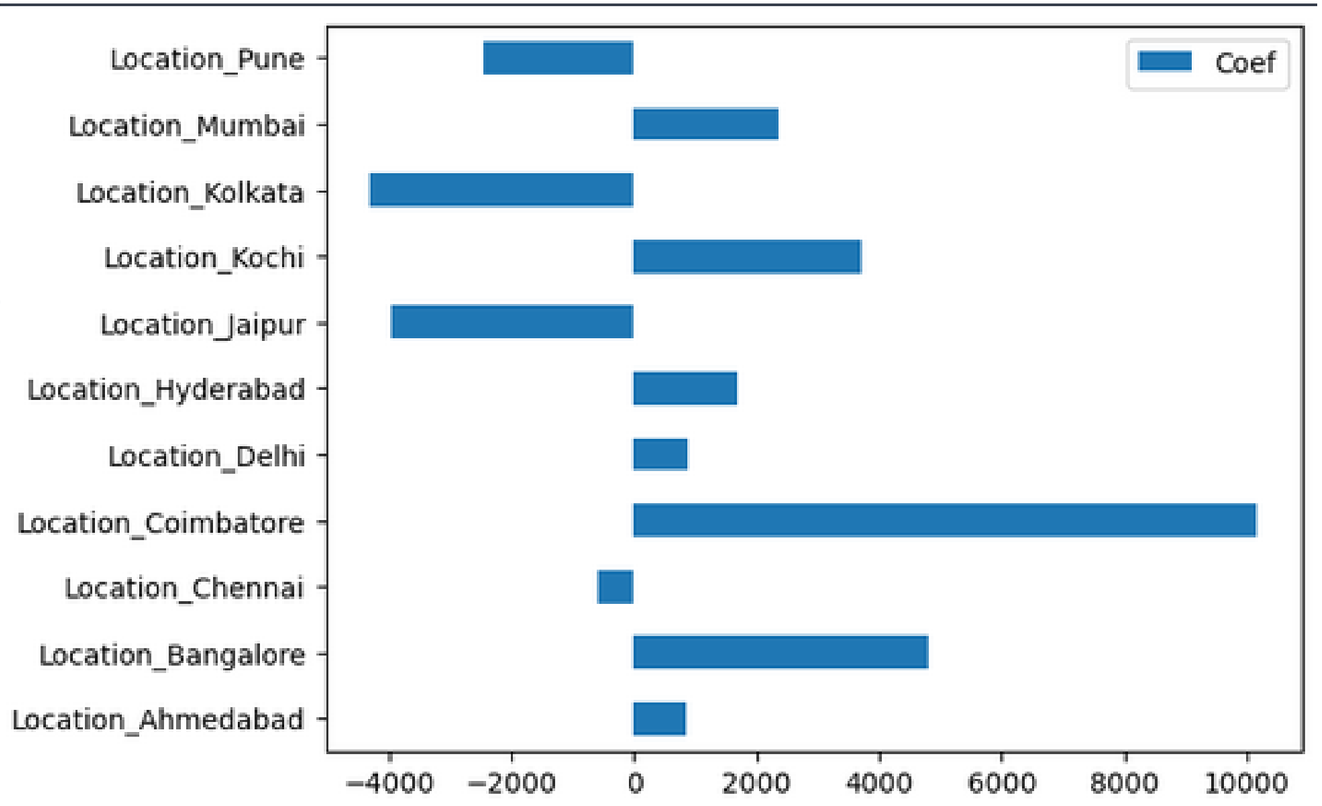
get_dummies를 사용하여 Location변수를 분리하였다.

다중선형회귀분석 모델의 설명변수로 Location만 넣고 실행한 결과 Brand별로 설명변수 중요도가 크게 달랐다.

이전 ANOVA를 통해 Location별로 가격에 주는 영향이 다르다는 결과와 부합하는 모습이다.

Mumbai, Kochi, Coimbator, Bangalor등의 경우 Price에 양의 방향으로 영향을 준다.

반면 Kolkata, jaipur등 의 경우 Price에 음의 방향으로 영향을 준다.



모델 생성

회귀분석

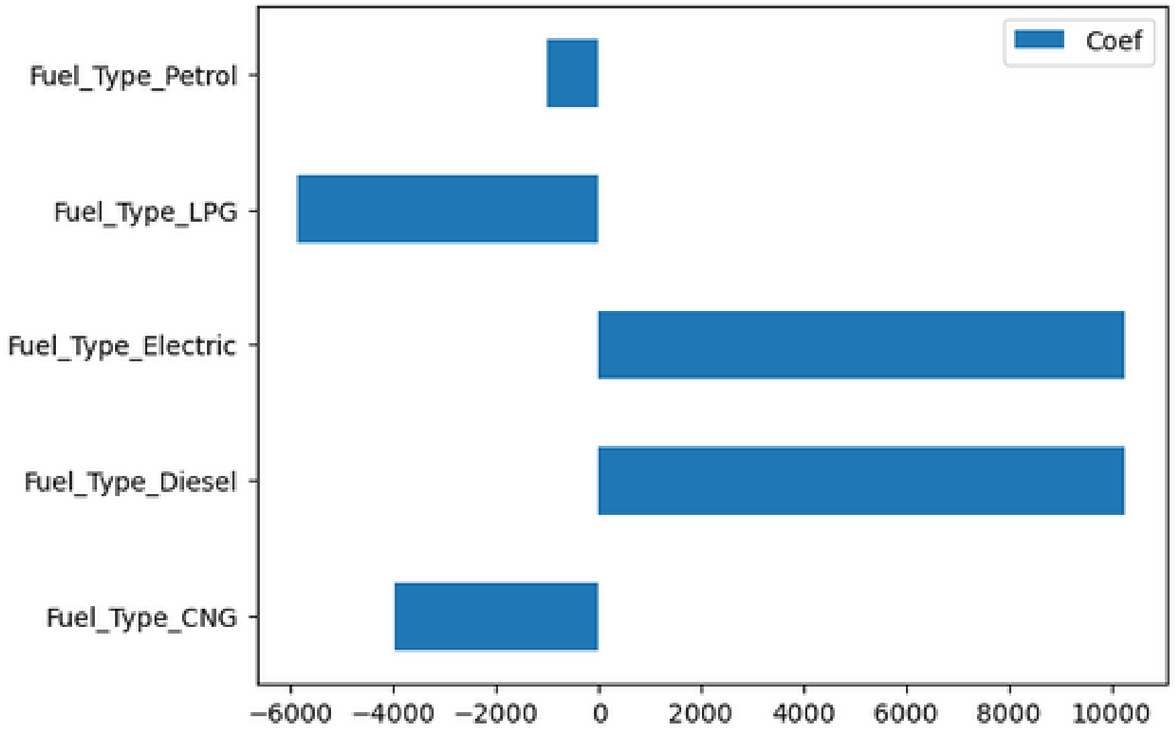
get_dummies를 사용하여 Fuel_Type 변수를 분리하였다.

다중선형회귀분석 모델의 설명변수로 Fuel_Type만 넣고 실행한 결과 Brand별로 설명변수 중요도가 크게 달랐다.

이전 ANOVA를 통해 Fuel_Type별로 가격에 주는 영향이 다르다는 결과와 부합하는 모습이다.

Electric, Diesel의 경우 Price에 양의 방향으로 영향을 준다.

반면 LPG, CNG 등 의 경우 Price에 음의 방향으로 영향을 준다.

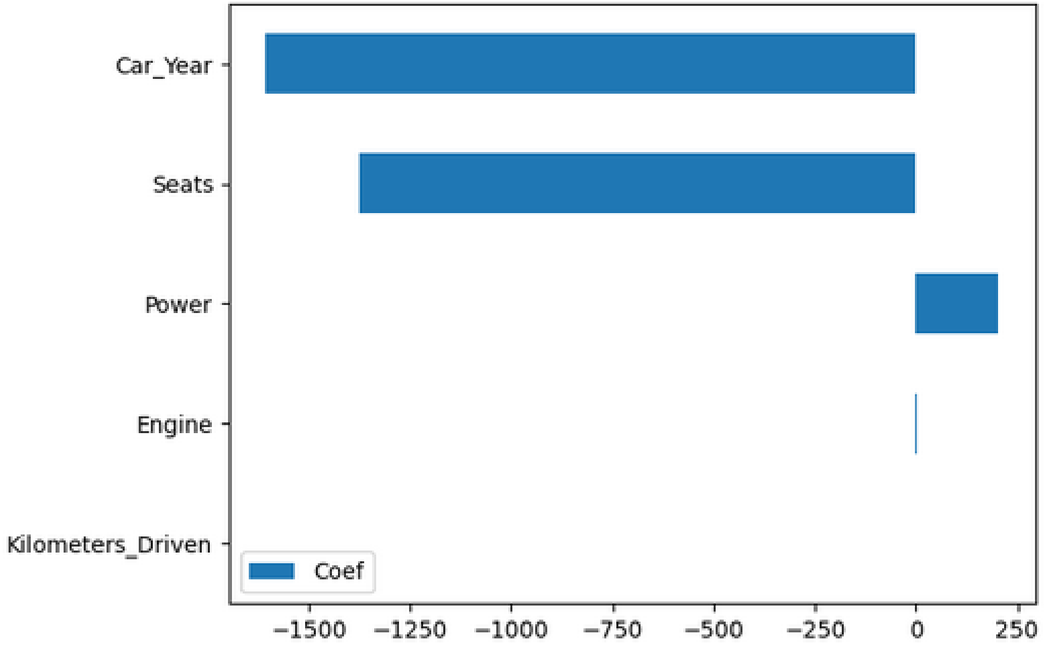


모델 생성

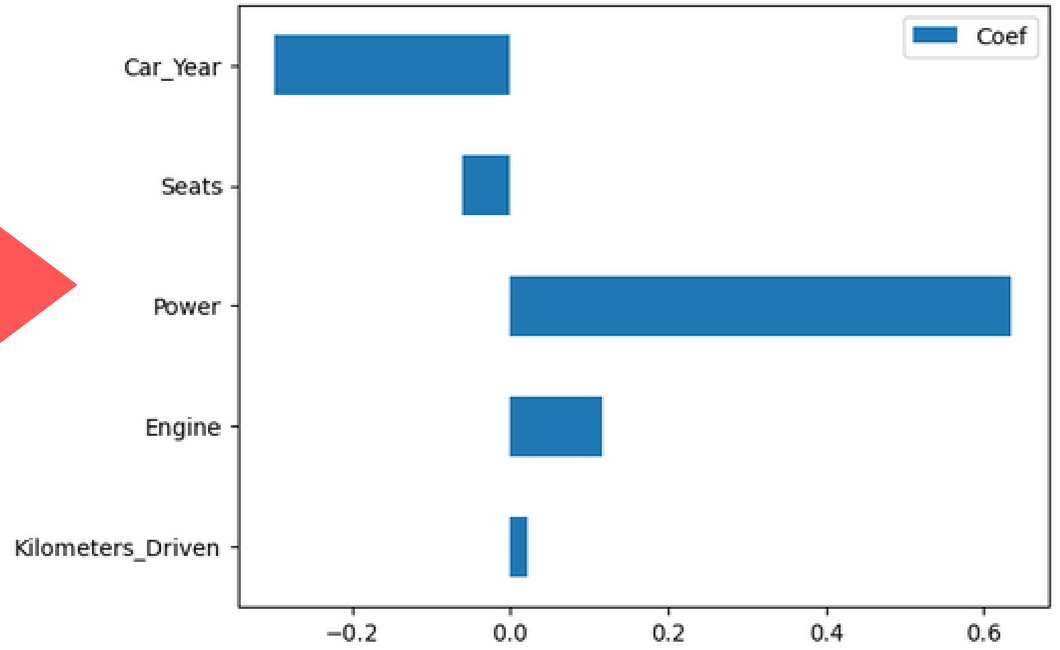
회귀분석

	variable	VIF
0	const	1.00
1	Kilometers_Driven	1.03
5	Car_Year	1.05
4	Seats	1.68
3	Power	5.06
2	Engine	6.09

주어진 모든 설명변수를 가지고 LinearRegression을 수행하면서 VIF가 너무 높은 변수들을 하나씩 제거한 결과 남은 변수들은 Kilometers_Driven, Car_Year, Seats, Power, Engine이다.
남은 변수들로 LinearRegression을 한 결과 모든 변수의 VIF지수가 모두 10 이하로 다중공선성이 해결되었다. 남은 변수들이 모두 수치형 변수들이라 데이터를 scale을 한 뒤 다시 비표준화 회귀계수 그래프를 확인하니 수치가 크게 변한 것을 알 수 있다. Power, Car_Year, Engine, Seats, Kilometers_Driven 순으로 Price에 영향을 많이 준다고 할 수 있다.

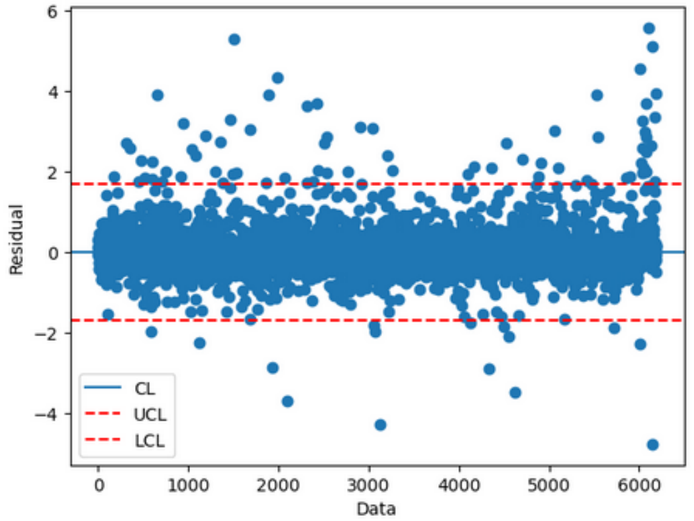


scale

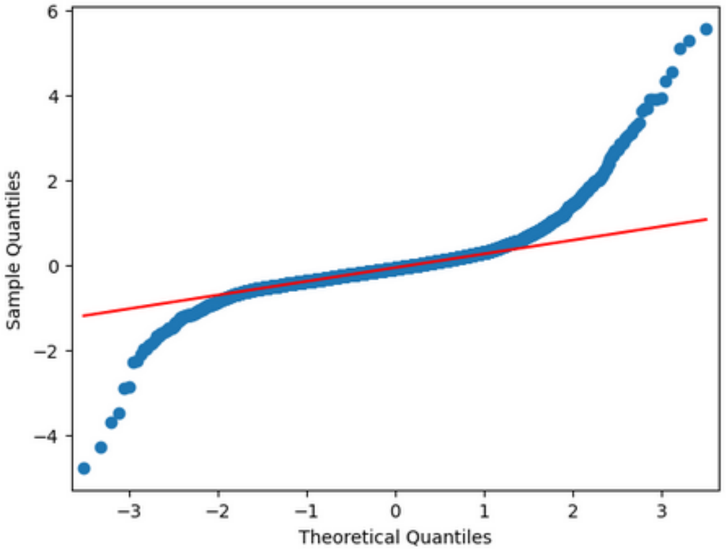


모델 생성

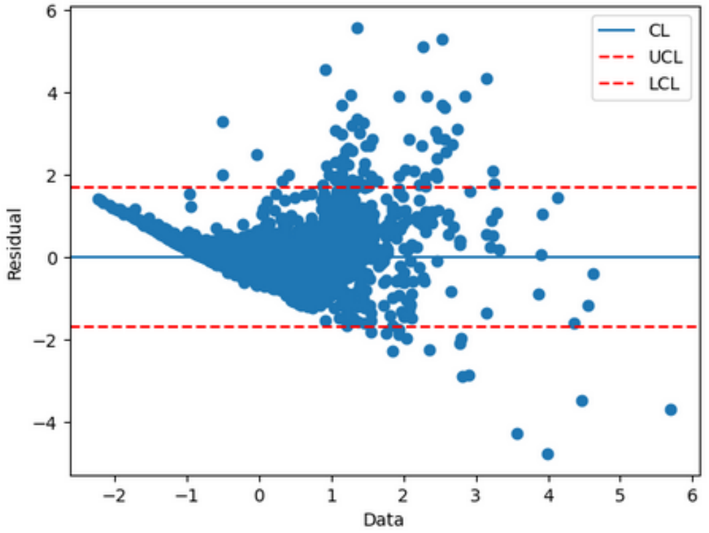
회귀분석



등분산성 검정



잔차 정규성 검정



잔차 독립성 검정

하지만 회귀분석에 필수적인 오차의 기본과정을 검토한 결과 문제 확인

잔차가 y=0을 기준으로 random하게 산포되어 있는 편이라 등분산성은 괜찮다고 할 수 있지만,

잔차의 정규성 검정에서 잔차가 정규분포 직선(적색선)에 벗어나는 부분이 많아서 잔차가 정규성을 띄지 않고

잔차의 독립성 검정에서 random하게 산포되어 있지 않아서 독립성을 만족하지 않는다.

따라서 이 모델을 사용하기에는 무리가 있다고 판단하였다.

모델 생성

의사결정나무

```
tree_uncustomized = DecisionTreeRegresso
tree_uncustomized.fit(df_train_x, df_tra

print("Score on training set: {:.3f}".fo
print("Score on test set: {:.3f}".format
```

```
Score on training set: 1.000
Score on test set: 0.760
```

```
tree_final = DecisionTreeRegressor(min_samples
tree_final.fit(df_train_x, df_train_y)

print("Score on training set: {:.3f}".format(t
print("Score on test set: {:.3f}".format(tree_
```

```
Score on training set: 0.792
Score on test set: 0.779
```

랜덤포레스트

```
rf_uncustomized = RandomForestRegressor(random
rf_uncustomized.fit(df_train_x, df_train_y)

print("Score on training set: {:.3f}".format(r
print("Score on test set: {:.3f}".format(rf_un
```

```
Score on training set: 0.981
Score on test set: 0.840
```

```
rf_final = RandomForestRegressor(min_samples_le
                                 max_depth = 15,
rf_final.fit(df_train_x, df_train_y)

print("Score on training set: {:.3f}".format(rf
print("Score on test set: {:.3f}".format(rf_fin
```

```
Score on training set: 0.884
Score on test set: 0.833
```

그라디언트 부스팅

```
gb_uncustomized = GradientBoostingRegressor
gb_uncustomized.fit(df_train_x, df_train_y)


print("Score on training set: {:.3f}".forma
print("Score on test set: {:.3f}".format(gb
```

```
Score on training set: 0.921
Score on test set: 0.828
```

```
gb_final = GradientBoostingRegressor(ran
                                     min_samp
gb_final.fit(df_train_x, df_train_y)

print("Score on training set: {:.3f}".fo
print("Score on test set: {:.3f}".format
```

```
Score on training set: 0.983
Score on test set: 0.843
```

의사결정나무, 랜덤포레스트, 그라디언트 부스팅 모두 기본옵션으로 모델을 생성했을 경우 모델 설명력이 train 데이터에 과대적합했다.

RandomizedSearchCV를 통해 모델의 parameter를 조정하여 과대적합 문제를 해결하였다.

모델 비교

	Feature	importance
3	Power	0.798
5	Car_Year	0.176
4	Seats	0.010
1	Mileage	0.008
2	Engine	0.005
23	Transmission_Automatic	0.003
0	Kilometers_Driven	0.000

	Feature	importance
3	Power	0.757
5	Car_Year	0.173
2	Engine	0.019
1	Mileage	0.019
0	Kilometers_Driven	0.014
6	Distance_Year	0.004
4	Seats	0.004
23	Transmission_Automatic	0.004
24	Transmission_Manual	0.003

	Feature	importance
3	Power	0.757
5	Car_Year	0.173
2	Engine	0.019
1	Mileage	0.019
0	Kilometers_Driven	0.014
6	Distance_Year	0.004
4	Seats	0.004
23	Transmission_Automatic	0.004
24	Transmission_Manual	0.003

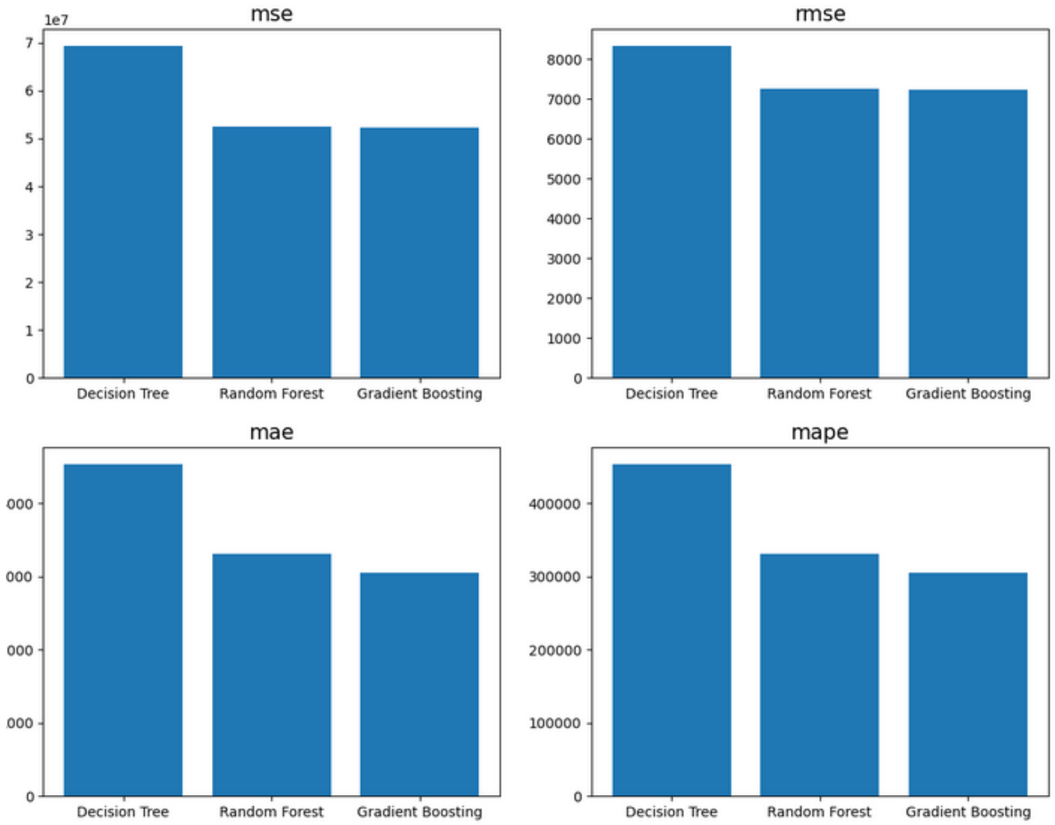
의사결정나무, 랜덤포레스트, 그라디언트 부스트 모두 상위 7개의 Feature이 일치했다.

일치하는 Feature는 다음과 같다.

Power, Car_Year, Engine, Mileage, Kilometers_Driven, Seats, Transmission_Automatic

특히 랜덤포레스트와 그라디언트 부스트는 일치하는 Feautre가 모두 동일한데 순서도 모두 동일하게 나왔다.

모델 비교



각 모델 별로

mean_squared_error

root_mean_squared_error

mean_absolute_error

mean_absolute_percentage_error

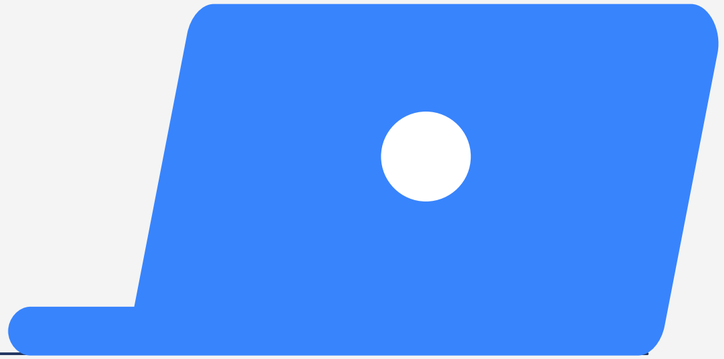
를 비교한 결과, 의사결정나무의 정확도가 제일 좋지 않고 그 다음으로 랜덤포레스트였으며 그라디언트팅부스팅이 다른 모델들보다 정확도가 가장 높은 것을 알 수 있었다.

따라서 그리디언트부스팅의 결과를 이용하는 것이 분석에 제일 바람직 하다고 판단하였다.

하지만 그라디언트부스팅과 랜덤포레스트의 결과가 동일했으므로 그 결과는 믿음직스럽다고 판단하였다.

랜덤포레스트와 그라디언트부스팅에서 나온 feature_importance 결과에 따르면,
중고차 판매 가격 예측 변수인 price에 가장 큰 영향을 끼치는 변수로는
마력(Power)이며, 그 다음으로는 현재연도 - 모델의 년도(Car_Year)이다.

또한, 주행거리(Kilometers_Driven)는 0.014의 importance score를 가지고,
배기량(Engine)과 회사에서 제공하는 표준주행거리(Mileage)는 0.019로 거의 유사한 값을 보이며,
좌석수(Seats)와 연식 대비 주행거리(Distance_Year)는 0.004으로 가장 낮은 importance score
를 보인다.



01

Power, Engine
마력, 배기량

Power, Engine이 높을수록
Price가 높다

02

Car_Year
현재연도 - 모델의 년도

Car_Year가 작을수록
Price가 높다

03

Seats
좌석 수

Seats가 작을수록
Price가 높다

04

Mileage
회사에서 제공하는 표준주행거리

Mileage가 낮을수록
Pricer가 높다

05

Distance_Year
연식 대비 주행거리

Distance_Year가 낮을수록
Price가 높다.

06

Transmission
변속기 종류

Automatic보다 Manual의
Price가 높다

최종 결론

중고차 가격을 효과적으로 예측할 수 있는 핵심영향인자는
마력의 크기와, 자동차의 연식이고
그 외에 추가적으로 고려해야할 인자는 주행거리이다.

하지만 브랜드, 거래 장소, 연료 종류, 소유권에 따른 차이도 고려하면 더욱 좋을 것이다.

중고차 가격은 판매자와 구매자의 매매 조건의 차이에 따라 큰 영향을 받으므로,
구체적인 시장 상황 및 구매자와 판매자의 입장에서도 고려해야 한다.