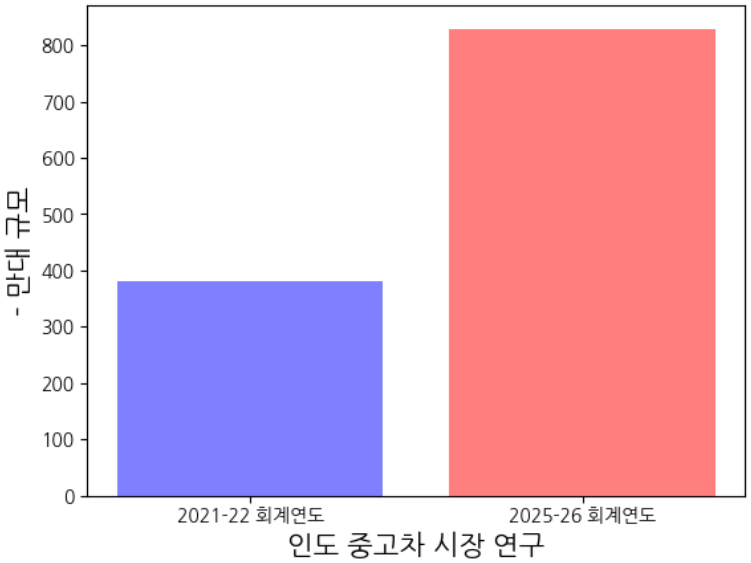
중고차 가격 예측모델 개발

분석 배경

현재 POS_Cars㈜는 인도의 중고차 시장을 진출을 목표로 하고 있다.
해당 업계는 최근 기술을 기반으로 한 서비스를 제공하는 스타트업들이 유입되어 시장을 점유하였고, 스타트업 업체들은 저렴한 가격의 부품과 효율적인 물류 서비스를 제공하고 있다.
인도의 온라인 플랫폼이 중고차 검색, 구매, 보험 가입에서 판매까지 이르는 모든 과정을 디지털화하는데 속도를 내고 있어 중고차 거래 과정을 크게 간소화하였다. 또한, 해당 업계는 최근 기술을 기반으로 한 서비스를 제공하는 스타트업들이 유입되어 시장을 점유하였고, 스타트업 업체들은 저렴한 가격의 부품과 효율적인 물류 서비스를 제공하고 있다.



프로스트&설리번社와 DWA社가 공동 발표한 인도 중고차 시장 연구' 보고서에 따르면 인도 중고차 시장은 2021-22회계연도에 380만 대에서 2025-26회계연도에는 820만대 규모로 성장할 것으로 전망

인도의 컨설팅업체인 Redseer는 보고서에서 온라인 플랫폼을 기반으로 한 중고차 서비스 및 수리 시장이 연평균 12% 성장하여, 2020년 80억 달러에서 2030년에는 250억의 시장 가치를 가질 것이라고 밝힘

목표 – 경쟁력 확보

[단독] 기아차, 인도 중고차 시장 진출 선언

국내의 대기업 또한 인도 중고차 시장에 진출하였다.

- 현재 시장 리더인 마루티 스즈키를 포함해 혼다인도, 마힌드라&마힌드라, 포드가 모두 중고차 사업을 병행하고 있다. 자매사인 현대자동차까지 H프로미스라는 브랜드로 중고차 사업을 진행하고 있다.
- 이에 따라 신생 진출기업이 경쟁력을 얻는 것은 매우 어려움이 있을 것으로 예상되어 진다.

매우 성장해나가는 중고차 시장에 진입하는 신규사업인 만큼,
경쟁 기업보다 효과적으로 중고차 판매 매출을 올리기 위해서

핵심 영향인자가 무엇이 있는지 파악하고 데이터 분석을 통한 빠른 중고차 가격예측

분석 계획

중고차 시세에 영향을 미치는 주요 요인

중고자동차의 가격에 영향을 미치는 내적 요소로는 연식, 주행거리, 사고유무, 차량상태, 옵션, 변속기의 종류, 색상, 사용 용도 등이 있다. 외적 요인인 소비자의 선호도나 계절적인 요인, 유행, 지역 등도 가격에 영향을 주는 사항이다. 하지만 중고차의 주요 수요요인은 나라마다 다른 특성을 지니고 있으므로 우리의 진출 시장인 '인도' 에서 중고차 가격의 주요 영향인자가 무엇인지 파악하도록 하겠다.

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

- 주어진 중고차 가격 데이터의 탐색적 분석 및 인도시장의 특성을 파악
- 중고차 가격의 일반적 요인을 기반으로 가설을 수립
- 데이터의 이상치와 결측치를 확인하여 데이터를 정제
- 잠재인자 해석 및 인도시장의 핵심 영향인자를 도출
- 복수의 예측 모델 개발 및 모델 평가 후 Best Model 선정
- 인도 중고차 시장의 진출 방향성 제시

결측치 처리 및 이상치 대체

```
df.isnull().sum()
```

```
Name                  0
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
dtype: int64
```

우리가 예측해야 할 목표변수 Price에 결측치 존재 확인
임의의 수를 넣어주면 목표변수의 해석방향이 달라질 것 같아서
결측치 행 제거하기로 결정

```
null = len(df_raw[df_raw['New_Price'].isnull()]) / len(df_raw)

print('New_Price의 결측치 비율:', null * 100)


# New_price 열은 결측치가 너무 많아서 설명력이 떨어질 것으로 예상된다.
# 열을 제외시키기로 결정하였다.
```

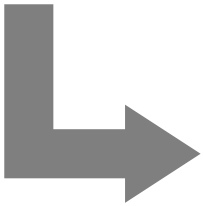
```
New_Price의 결측치 비율: 83.69070825211176
```

New_price 열은 결측치가 너무 많아서 적은 수의 데이터로 New_price의 특성 설명해주기 어려울 것으로 판단되어 열을 제거

- Power에 null 문자 들어간 행들은 Power의 평균으로 결측치 대체해주었다.
- Kilometers_Driven, Price 열에서 유독 큰 값을 가진 행 삭제 (이상치)

데이터 정제

	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

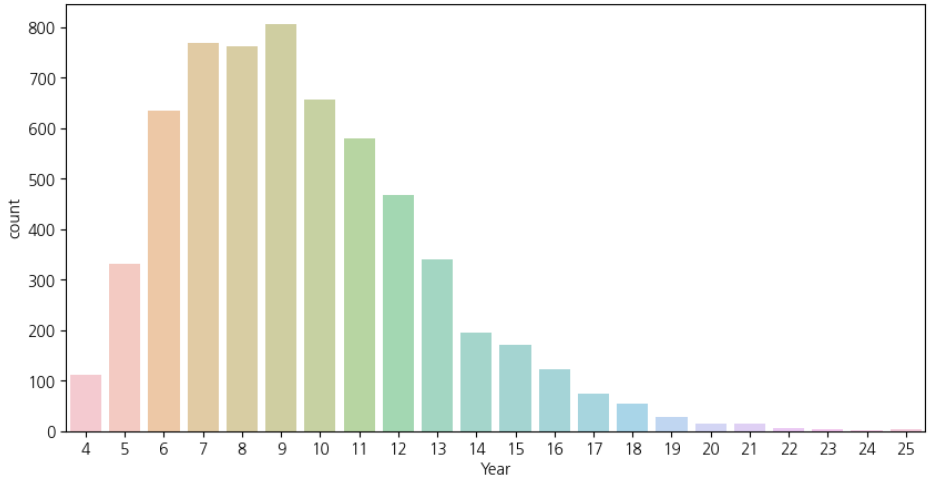


	Name	Location	Price	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats
0	Maruti	Mumbai	2682.68	13	72000	CNG	Manual	1	26.60	998.0	58.16	5.0
1	Hyundai	Pune	19162.00	8	41000	Diesel	Manual	1	19.67	1582.0	126.20	5.0
2	Honda	Chennai	6898.32	12	46000	Petrol	Manual	1	18.20	1199.0	88.70	5.0
3	Maruti	Chennai	9197.76	11	87000	Diesel	Manual	1	20.77	1248.0	88.76	7.0
4	Audi	Coimbatore	27194.71	10	40670	Diesel	Automatic	2	15.20	1968.0	140.80	5.0
5	Hyundai	Hyderabad	3602.46	11	75000	LPG	Manual	1	21.10	814.0	55.20	5.0
6	Nissan	Jaipur	5365.36	10	86999	Diesel	Manual	1	23.08	1461.0	63.10	5.0
7	Toyota	Mumbai	26826.80	7	36000	Diesel	Automatic	1	11.36	2755.0	171.50	8.0
8	Volkswagen	Pune	7971.39	10	64430	Diesel	Manual	1	20.54	1598.0	103.60	5.0
9	Tata	Chennai	2989.27	11	65932	Diesel	Manual	2	22.30	1248.0	74.00	5.0

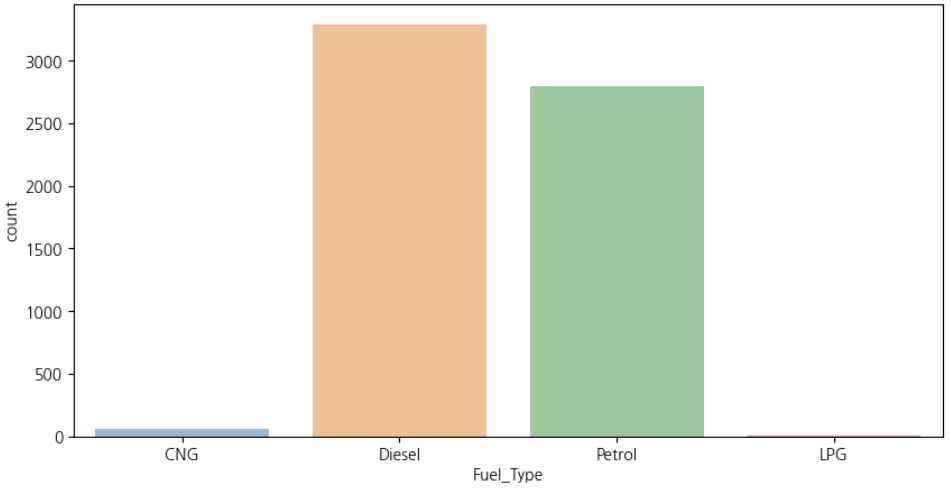
Name, Mileage, Engine, Power -> 단위를 잘라줌
Name 열을 맨 앞의 Brand만 남기도록 데이터 변경

Year -> 현재년도 - 생산년도 = 차 연식 열 생성
Owner_Type -> First : 1, Second : 2, Third : 3, Fourth & Above : 4

그래프 분석

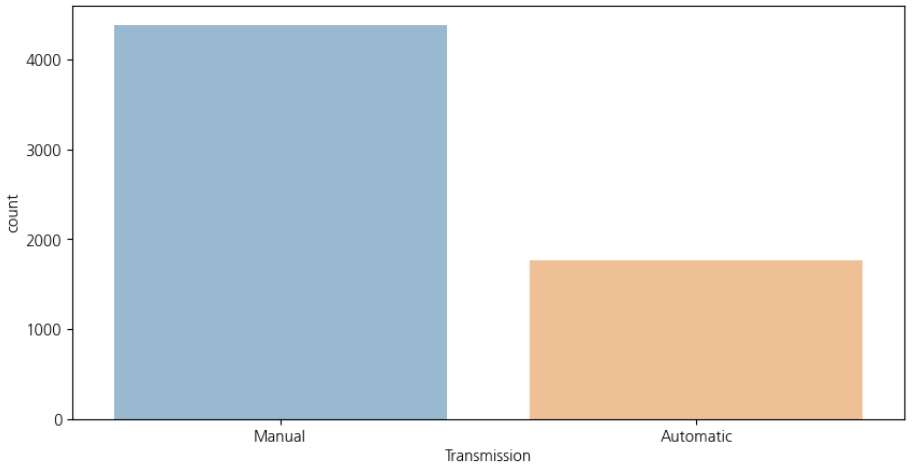


연식이 10년을 넘어갈수록
거래수가 줄어들고 있다.
소비자들은 오래된 연식의 자동차
선호하지 않는다는 걸 예측가능
-> 연식이 목표변수에도 영향을 많이
끼칠 것이라고 가설 설정 가능



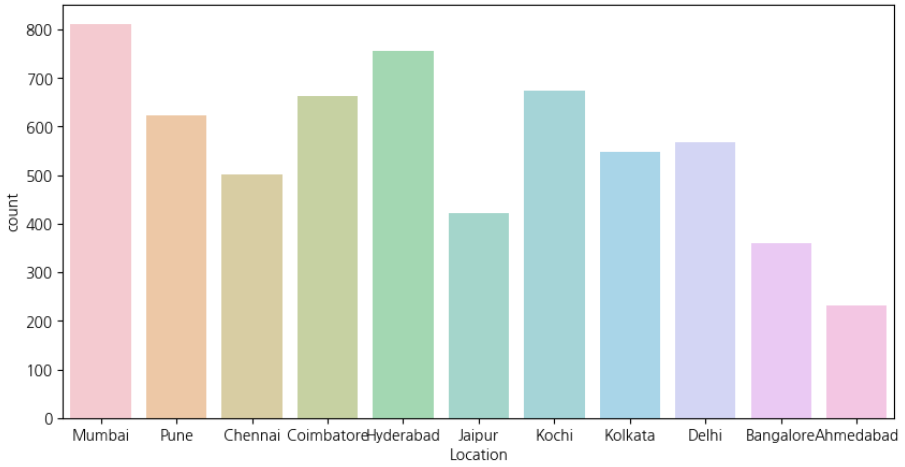
4가지 연료 타입 중에서
Desel과 Petrol의 가격 데이터가
많은 것을 확인
-> 두 집단의 가격 차이가 있을 것
이라고 가설 설정 가능

그래프 분석



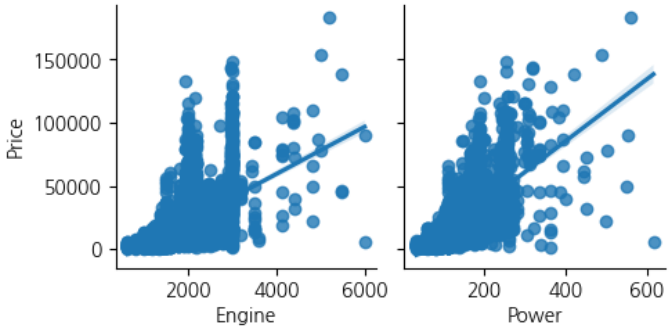
자동차의 사용 변속기의 종류 = Transmission의 경우 Manual이 많은 것을 확인

-> 인도 중고차 시장의 선호로 으 생각해볼 수 있음



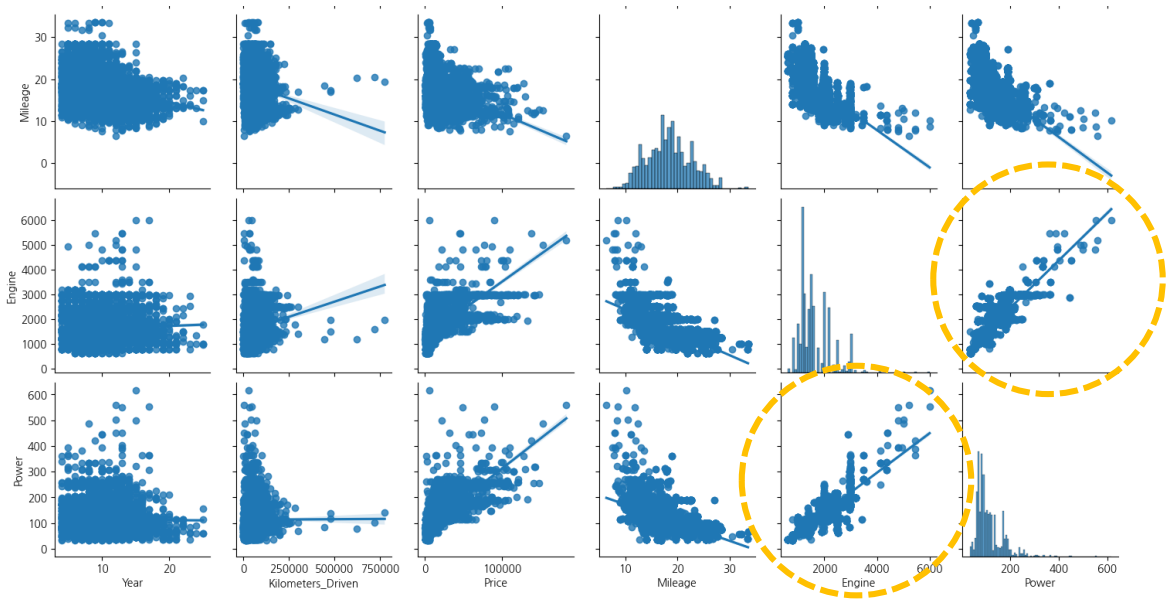
지역마다 큰 차이는 없지만 Jaipur, Bangalore, Ahmedabad 가 낮은 거래 수를 보이는 것으로 확인
분석 후에 특정 지역만 따로 뽑아 볼 필요성 존재함

탐색적 분석



Price와 Power, Engine이 선형성을 보인다.
Power와 Engine이 영향력 있을 것 이라고 예측 가능

탐색적 분석



하지만, Engine과 Power가 서로 상관관계가 있는 것을 확인
다중공선성 처리의 필요성 존재

가설 설정

연식
가설 : 자동차 연식이 가격 측정에 영향을 미칠 것이다.

자동차 연료 (Fuel_Type)
가설 : 자동차 종의 연료에 따른 가격 평균에 차이가 있을 것이다.

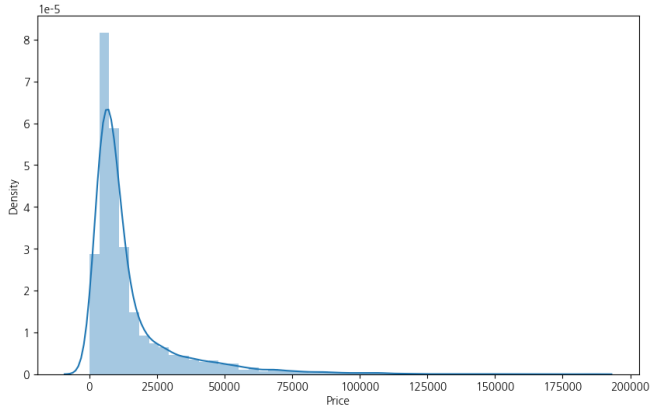
자동차 브랜드 별 (Name)
가설 : 자동차 브랜드가 가격 측정에 영향을 미칠 것이다.

자동차의 변속기 종류 (Transmission)
가설 : 변속기 종류가 가격 측정에 영향을 미칠 것이다.

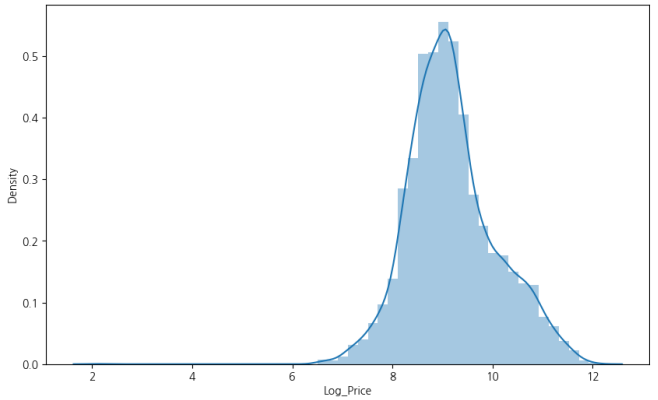
정규성 검정 및 Scaling 진행

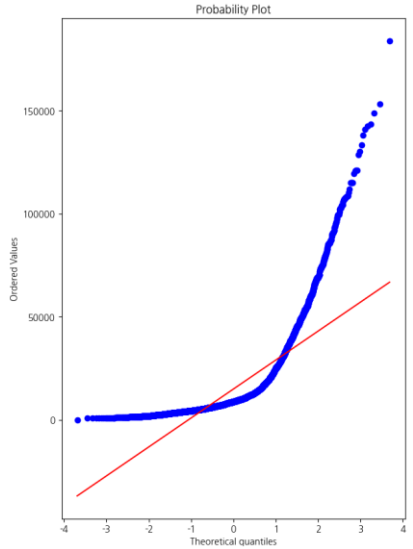
설명변수 : 설명변수 간의 데이터 분포를 맞춰주기 위해 Data scale 필요
목표변수 : 목표변수가 정규성을 충족하지 못 해 log화 필요
자동차 연료 (Fuel Type) , 자동차 변속기 종류(Transmission) : ANOVA 분석 (군집간의 평균 차이) 필요

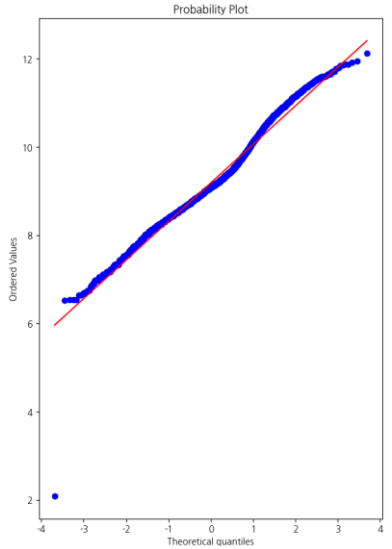
탐색적 분석



목표변수 Price
Log화 진행







목표변수가
훨씬 정규분포를
따르게 되었다.

가설 검정

ANOVA 분산검정

자동차 연료
대립가설 : 자동차 종의 연료에 따른 가격 평균에 차이가 있을 것이다.

```
statistic, p = stats.shapiro(disel)
print('disel statistic = {}, p-value ={}'.format(statistic, p))

# 귀무가설 - 데이터는 정규성을 지니고 있다.
# p-value가 0.05보다 작으므로 귀무가설을 기각, 즉 데이터는 정규성을 만족한다고 볼 수 없습니다
```

```
disel statistic = 0.9854835867881775, p-value =7.103975165935049e-18
```

```
statistic, p = stats.shapiro(petrol)
print('disel statistic = {}, p-value ={}'.format(statistic, p))

# p-value가 0.05보다 작으므로 귀무가설을 기각, 즉 데이터는 정규성을 만족한다고 볼 수 없습니다
```

```
disel statistic = 0.9639075994491577, p-value =7.000044140437739e-26
```

- 연료 타입에 따른 분산 분석을 하려했으나, 데이터가 정규성을 만족하지 않아서 분산 분석 미실시

다중회귀분석

OLS Regression Results

Dep. Variable:	Log_Price	R-squared:	0.810
Model:	OLS	Adj. R-squared:	0.809
Method:	Least Squares	F-statistic:	2611.
Date:	Mon, 06 Mar 2023	Prob (F-statistic):	0.00
Time:	04:05:07	Log-Likelihood:	-1711.6
No. Observations:	3691	AIC:	3437.
Df Residuals:	3684	BIC:	3481.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.7921	0.089	99.210	0.000	8.618	8.966
Year	-0.1358	0.002	-56.703	0.000	-0.140	-0.131
Kilometers_Driven	-2.554e-07	2.04e-07	-1.255	0.210	-6.55e-07	1.44e-07
Mileage	0.0026	0.002	1.225	0.221	-0.002	0.007
Engine	0.0004	2.72e-05	14.141	0.000	0.000	0.000
Power	0.0083	0.000	30.065	0.000	0.008	0.009
Seats	0.0150	0.010	1.468	0.142	-0.005	0.035

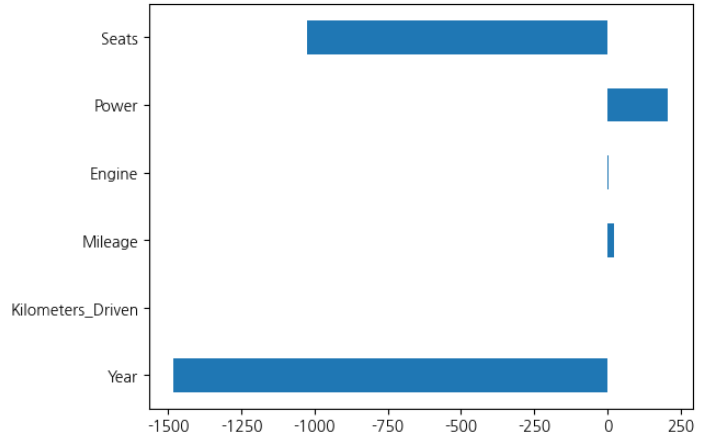
Omnibus:	1729.029	Durbin-Watson:	2.049
Prob(Omnibus):	0.000	Jarque-Bera (JB):	77668.664
Skew:	-1.504	Prob(JB):	0.00
Kurtosis:	25.271	Cond. No.	9.48e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.48e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

[1] 표준 오차는 오차의 공분산 행렬이 올바르게 지정되었다고 가정합니다.

[2] 조건 번호는 9.48e+05로 큽니다. 이는 다음이 있음을 나타낼 수 있습니다 강력한 다중 공선성 또는 기타 수치 문제.

--> 데이터 스케일링이 필요한 것을 확인할 수 있다.



[데이터 스케일링 전]

변수 중요도로 Year, Seats 열이 음의 영향력을 지니고 있다.

다중회귀분석

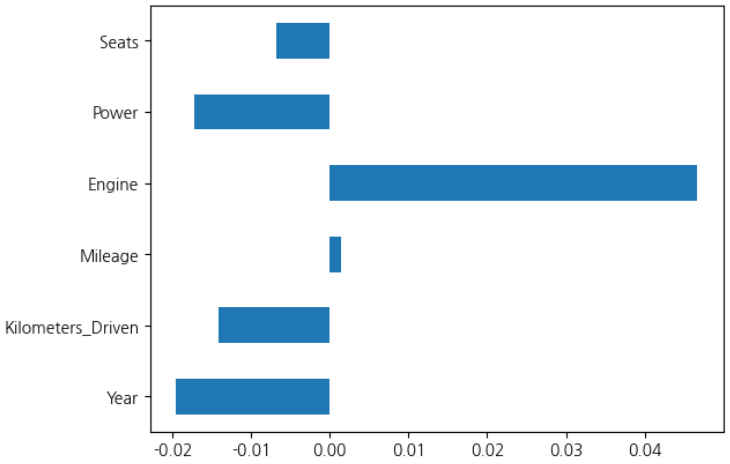
OLS Regression Results

Dep. Variable:	Log_Price	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	1.906
Date:	Mon, 06 Mar 2023	Prob (F-statistic):	0.0759
Time:	04:05:08	Log-Likelihood:	-7667.7
No. Observations:	5989	AIC:	1.535e+04
Df Residuals:	5982	BIC:	1.540e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.1650	0.011	814.143	0.000	9.143	9.187
Year	-0.0195	0.014	-1.416	0.157	-0.047	0.008
Kilometers_Driven	-0.0141	0.013	-1.064	0.287	-0.040	0.012
Mileage	0.0014	0.016	0.089	0.929	-0.030	0.033
Engine	0.0467	0.029	1.591	0.112	-0.011	0.104
Power	-0.0172	0.027	-0.647	0.517	-0.069	0.035
Seats	-0.0067	0.015	-0.454	0.650	-0.036	0.022

Omnibus:	163.615	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	176.682
Skew:	0.412	Prob(JB):	4.31e-39
Kurtosis:	3.167	Cond. No.	5.60

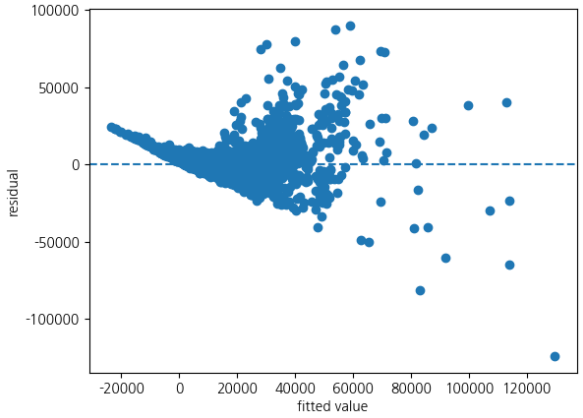
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

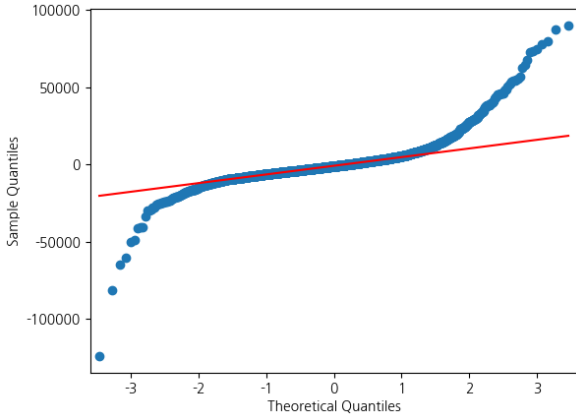


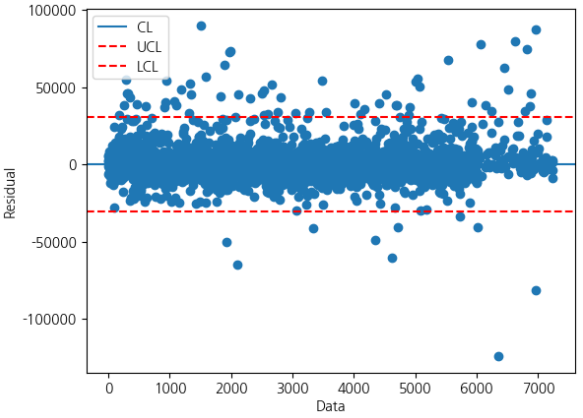
[데이터 스케일링 후]

설명변수를 스케일링해주자 모델의 설명력이 현저히 떨어졌음을 확인
그러나 변수 중요도를 확인해보았을때, 여전히 year(연식)열이 중요도가 큰 것을 확인할 수 있다.
주요인자로 year 결정, Engine열도 주요인자로 주의 깊게 확인할 필요성 존재한다.

다중회귀분석







등분산성

정규성

잔차독립성

스케일링 후에 만들어진 모델이 회귀분석의 전제조건을 충족하지 못하는 것을 확인 할 수 있다.
해당 모델이 유의미한지 검토가 필요하다.

의사결정나무

최종모델 파라미터

```
tree_final = DecisionTreeRegressor(min_samples_leaf = 8,
                                   min_samples_split = 24, max_depth = 4,
                                   random_state = 1234)
```

변수 중요도

	Feature	Importance
4	Power	0.72721
0	Year	0.24448
3	Engine	0.02496
6	Name_Audi	0.00335

Score

```
Score on training set: 0.896393
Score on test set: 0.840753
```

가장 먼저 노드를 분할하는 변수는 Power, Year, Engine
- 주요인자로 도출

랜덤포레스트

최종모델 파라미터

```
rf_final = RandomForestRegressor(random_state = 1234,
           n_estimators = 50, min_samples_leaf = 30,
           max_depth = 5)
```

변수 중요도

	Feature	Importance
4	Power	0.70113
0	Year	0.25441
3	Engine	0.03733
48	Transmission_Manual	0.00213
1	Kilometers_Driven	0.00163

Score

```
Score on training set: 0.870
Score on test set: 0.832
```

가장 먼저 노드를 분할하는 변수는 Power, Year, Engine
- 주요인자로 도출

그래디언트부스팅

최종모델 파라미터

```
gb_final = GradientBoostingRegressor(random_state = 1234,
           n_estimators = 40, min_samples_leaf = 14,
           min_samples_split = 34, max_depth = 5,
           learning_rate = 0.3)
```

변수 중요도

	Feature	Importance
4	Power	0.68838
0	Year	0.15276
3	Engine	0.04131
2	Mileage	0.03106
1	Kilometers_Driven	0.03059
20	Name_Land	0.01006
48	Transmission_Manual	0.00810

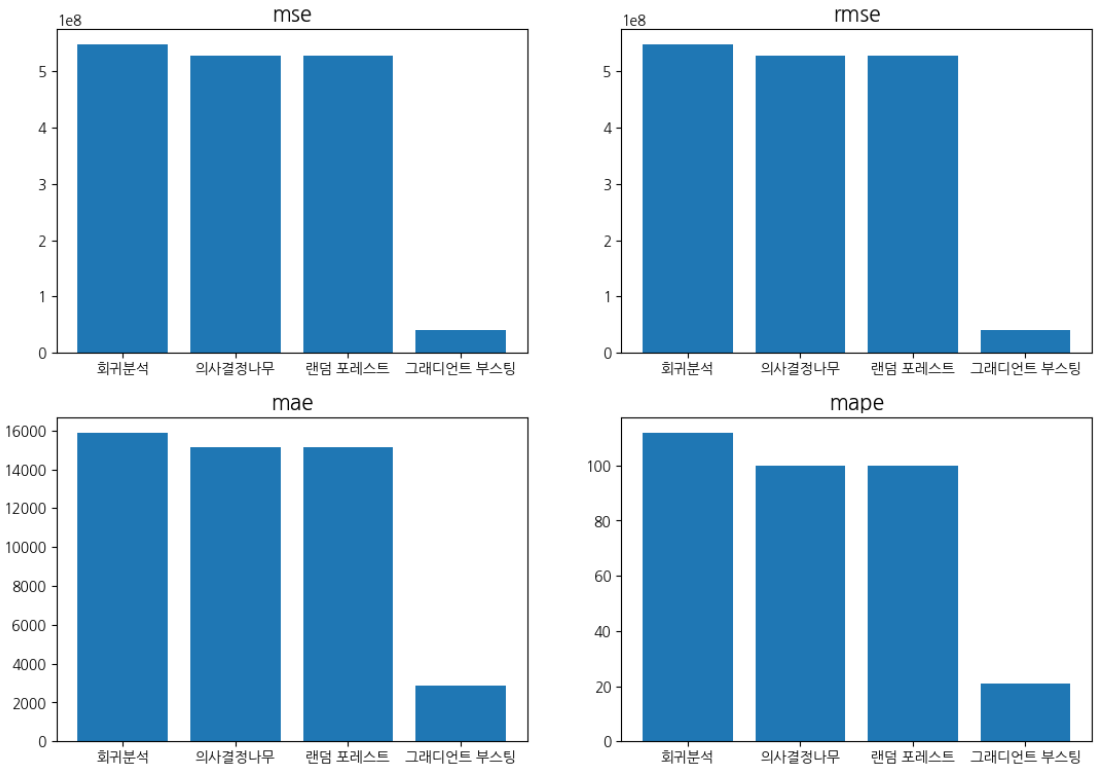
Score

```
Score on training set: 0.942
Score on test set: 0.865
```

가장 먼저 노드를 분할하는 변수는 Power, Year, Engine
- 주요인자로 도출

최종 모델 비교

최종모델 파라미터



평가지표 MAE는
Mean (평균) Absolute (절대값)
Error (잔차) 라는 의미로 잔차값
(예측값과 실제값 간의 차이. 즉,
모델이 예측을 얼마나 틀렸는가)
을 절대값 취하고 평균 낸 것

모델들의 예측성이 좋지 않다는 것을 알 수 있다.

결론

연식 <- Year
가설 : 자동차 연식이 가격 측정에 영향을 미칠 것이다.

자동차 연료 (Fuel_Type) <- 가설 기각
가설 : 자동차 종의 연료에 따른 가격 평균에 차이가 있을 것이다.

자동차 브랜드 별 (Name) <- 가설 기각
가설 : 자동차 브랜드가 가격 측정에 영향을 미칠 것이다.

자동차의 변속기 종류 (Transmission) <- 가설 기각
가설 : 변속기 종류가 가격 측정에 영향을 미칠 것이다.

모델 분석을 통해서 변수 중요도가 높은 변수
= Power, Year, Engine이 존재한다.

→

기존에 생각하지 못 했던
Power와 Engine 변수에
대해 알아볼 필요성 존재

개선안 제시

이번 유효한 변수로 선택한 Year, Power, Engine은 중고차의 가격을 결정 짓는 당연한 요인이라고 관념적으로 알 수 있는 변수이다. 이에 따라서 다음 데이터 분석을 진행한다면, 해당 요인을 제외하고 분석을 해봐도 좋을 것으로 예상된다.

- 우리나라 중고차 시장에서 '연비' Mileage의 경우, 큰 영향요인으로 여겨지지 않고 있다. 하지만 그래디언트 부스팅 모델에서 4번째 유효변수로 mileage가 존재한다. 이는 인도시장의 특성으로도 생각해볼 가치가 있다. 이전의 주요인자들이 모두 자동차의 '성능' 관련 변수들이었기 때문에 시장 진출시, 특정 지역과 특정 고급모델에 집중하는 것보다 자동차의 성능에 집중하여 중고차제품을 판매한다면 매출증대를 기대해볼 수 있을 것이다.

통찰

데이터에 대해서 내가 세운 가설들이 대부분 맞지 않다는 것을 깨닫게 되었다. 그리고 실제 데이터에서는 전처리와 정제가 나의 예상보다 중요하다는 것을 알게 되었다. 데이터의 단순한 선형성으로는 목표변수를 예측하기 어렵다는 것도 알게 되었다.

지역마다 목표변수도 변동이 있을 것이라고 예상이 되어지나, 분석 방법에 대해서 미흡함을 느꼈다. 데이터 검정에 정규성, 독립성, 분산성 조건을 맞추는 것에 대해서 더욱 공부 해야함을 느꼈다.

가설 검증에 있어서도 도메인 지식이 필요하다는 것을 알게 됐다. 그리고 분석을 함에 있어서 가설을 최대한 많이 세우고 조금씩 수정해보면서 데이터에 대해 이해해 나갈 수 있도록 하는 방향성을 지녀야할 것 같다.