중고차 예측 Project

임승호

TIMELINE

과제정의

배경

분석하고자 하는 방향

예상 목표

분석계획

데이터 확인

데이터 전처리

가설 검정

모델링

데이터 현황

데이터 확인

기술 통계량 확인

이상치 확인

결측치 확인

탐색적 분석

데이터 시각화

가설 검정

상관분석

모델링 및 평가

Regression

Decision Tree

RandomForest

Gradient Boosting

결론

결과 및 결론

종합결과

핵심인자 정리

과제정의

인도의 신차 판매가 주춤

POS_Cars가 중고차 시장에
뛰어들 준비를 하고 있다

중고차 시장에 뛰어들어 경쟁력과
수익성을 향상 시켜야 한다

중고차시장이 커지고 있다

고객들은 차량의 특성과 상태가
좋은 중고차를 선호한다

따라서 POS_Cars는 중고차 가격을
효과적으로 예측할 수 있는
가격예측 모델을 개발하려한다

분석계획

데이터 확인

결측치 확인
기술통계량 확인
이상치 확인

데이터 전처리

이상치 처리
결측치 처리
파생변수 생성

가설 검정

범주형 설명변수와 목표변수간 분포확인
연속형 설명변수와 목표변수간 분포 확인
설명변수와 목표변수간 차이 검정

모델링

다중선형회귀분석
의사결정나무
랜덤포레스트
그래디언트 부스팅
모델 평가

데이터 현황

#	Column	Non-Null Count	Dtype
0	Name	7253 non-null	object
1	Location	7253 non-null	object
2	Price	6200 non-null	float64
3	Year	7253 non-null	int64
4	Kilometers_Driven	7253 non-null	int64
5	Fuel_Type	7253 non-null	object
6	Transmission	7253 non-null	object
7	Owner_Type	7253 non-null	object
8	Mileage	7251 non-null	object
9	Engine	7207 non-null	object
10	Power	7207 non-null	object
11	Seats	7200 non-null	float64
12	New_Price	1006 non-null	object

목표 변수 : Price

범주형 설명 변수 : Name, Location, Transmission Owner_Type, Mileage, Engine, Power, New_price

연속형 설명변수 : Year, Kilometers_Driven, Seats

⬇

하지만 Mileage, Engine, Power는 원래

연속형 변수이지만,

단위때문에 object타입으로 명시되었다.

단위 변환

```
df['Mileage'] = df['Mileage'].str.replace('kmpl','').astype('float64')
df['Engine'] = df['Engine'].str.replace('CC','').astype('float64')
df['Power'] = df['Power'].str.replace('bhp','')
```

현재 Power 변수가 타입이 바뀌지 않는 문제가 생김

그래서 unique() 함수를 이용하여 값들을 확인해보니

"null "이 str형태로 저장되어 있었다

```
df['Mileage'] = df['Mileage'].str.replace('kmpl','').astype('float64')
df['Engine'] = df['Engine'].str.replace('CC','').astype('float64')
df['Power'] = df['Power'].str.replace('bhp','')
```

```
df[df['Power']=='null '].head(5)
```

	Name	Location	Price	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price
76	Ford Fiesta 1.4 SXi TDCi	Jaipur	3065.92	2008	111111	Diesel	Manual	First	17.8	1399.0	null	5.0	NaN
79	Hyundai Santro Xing XL	Hyderabad	1992.85	2005	87591	Petrol	Manual	First	0.0	1086.0	null	5.0	NaN
89	Hyundai Santro Xing XO	Hyderabad	3219.22	2007	73745	Petrol	Manual	First	17.0	1086.0	null	5.0	NaN
120	Hyundai Santro Xing XL eRLX Euro III	Mumbai	1303.02	2005	102000	Petrol	Manual	Second	17.0	1086.0	null	5.0	NaN
143	Hyundai Santro Xing XO eRLX Euro II	Kochi	2560.04	2008	80759	Petrol	Manual	Third	17.0	1086.0	null	5.0	NaN

파생변수 생성

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

Name의 첫 단어만 뽑아 Brand라는 파생변수 생성

그 다음 Name 이라는 변수는 unique하기 때문에 drop시킴

그리고 현재 날짜를 기준으로 Year를 뺀 연식이라는 파생변수를

만드려 했지만, 회귀절편에 유의미한 영향을 주지 않을 것으로

판단하여 파생변수로 사용하지 않고 Year 변수 그대로 사용

이상치 확인

	Price	Year	Kilometers_Driven	Mileage	Engine	Power	Seats
count	6200.000000	6200.000000	6.200000e+03	6198.000000	6164.000000	6164.000000	6158.000000
mean	14912.514750	2013.434194	5.815738e+04	18.183448	1620.003731	111.346067	5.278500
std	17674.318464	3.271969	9.010627e+04	4.577602	601.493220	55.699437	0.808371
min	7.080000	1998.000000	1.710000e+02	0.000000	72.000000	0.000000	0.000000
25%	5365.360000	2012.000000	3.300000e+04	15.260000	1198.000000	74.000000	5.000000
50%	8814.520000	2014.000000	5.251450e+04	18.200000	1493.000000	92.350000	5.000000
75%	15869.972500	2016.000000	7.227750e+04	21.100000	1979.500000	138.100000	5.000000
max	245273.600000	2019.000000	6.500000e+06	33.540000	5998.000000	616.000000	10.000000

Kilometers_Driven

max

650만 km

50만 km이상 탄 차는 대부분 폐차를 한다

Mileage

min

0

연비가 0일 수 없다

Price

min

7만원

자동차 가격이 7만원 일 수 없다

Seats

min

0

자동차 좌석 수가 0일 수 없다

결측치 확인

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

➡

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

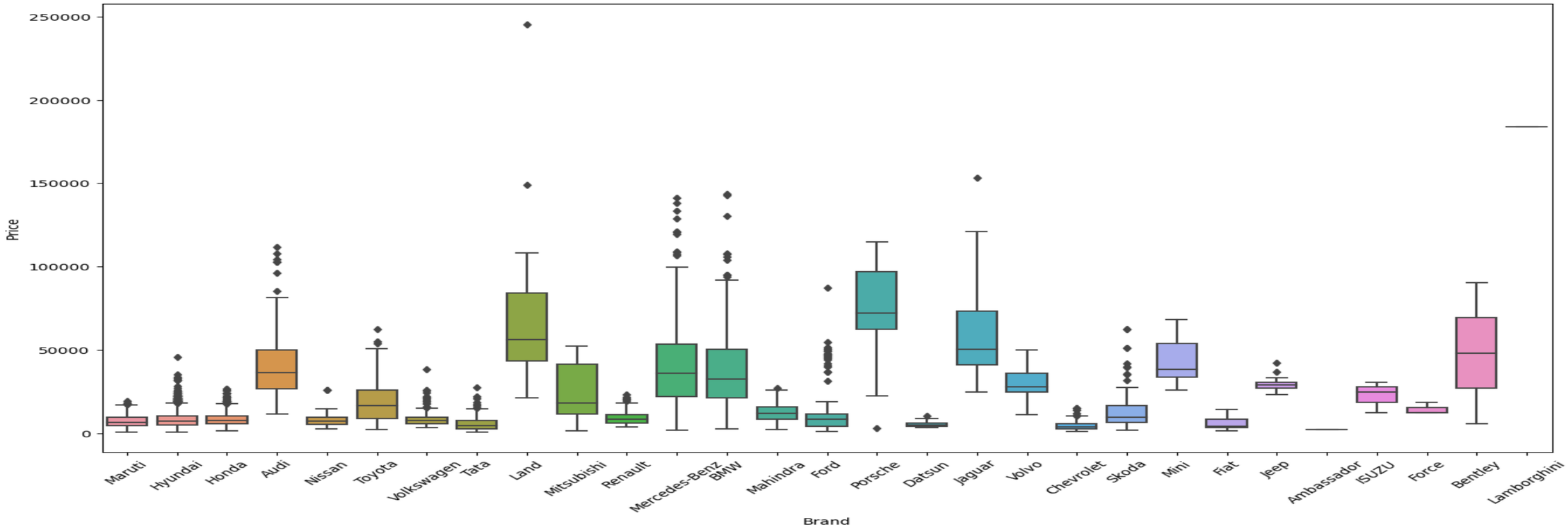
목표변수 Price의 결측치는 제외

Mileage, Engine, Power, Seated 모두
제조사(Brand) 평균별로 결측치를 채움

※결측치보다 이상치먼저 제거한 이유※

평균은 이상치에 민감하기 때문에 결측치를 채우고
이상치를 제거하면 값차이가 커지게된다

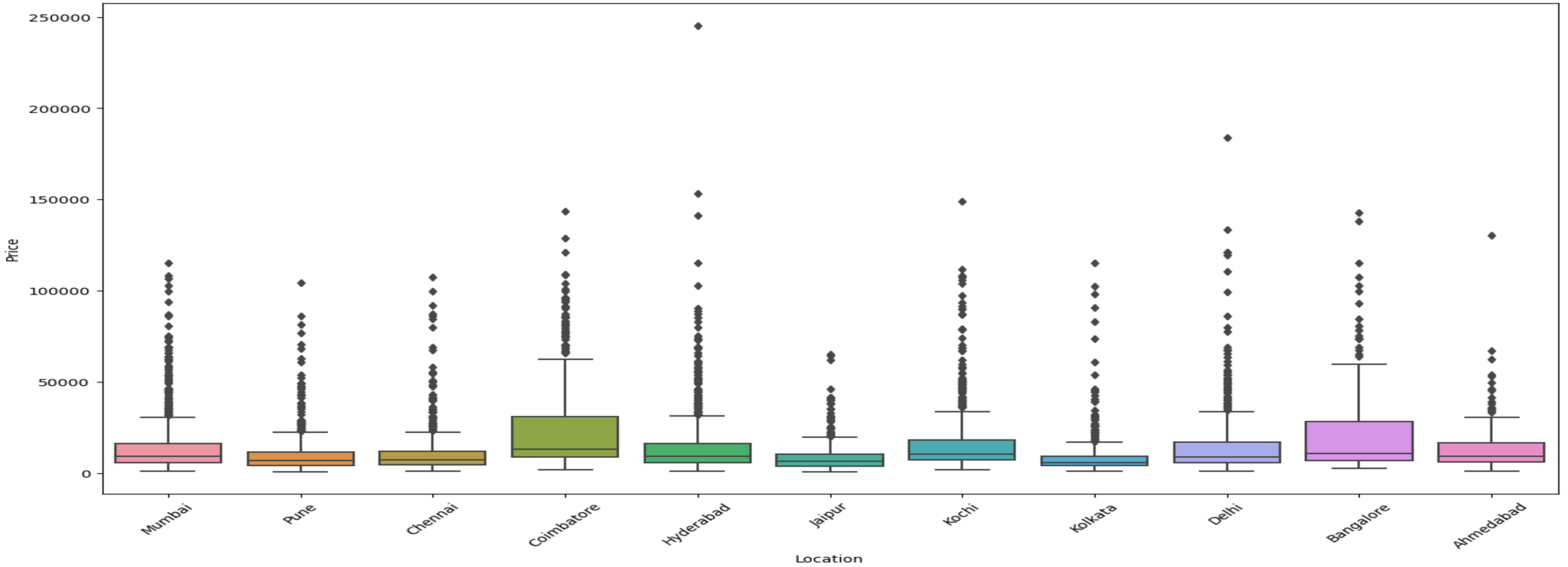
탐색적 분석



목표변수(Price)와 설명변수(Brand)의 분포확인

자동차 브랜드별로 가격대가 다른 것을 알 수 있다

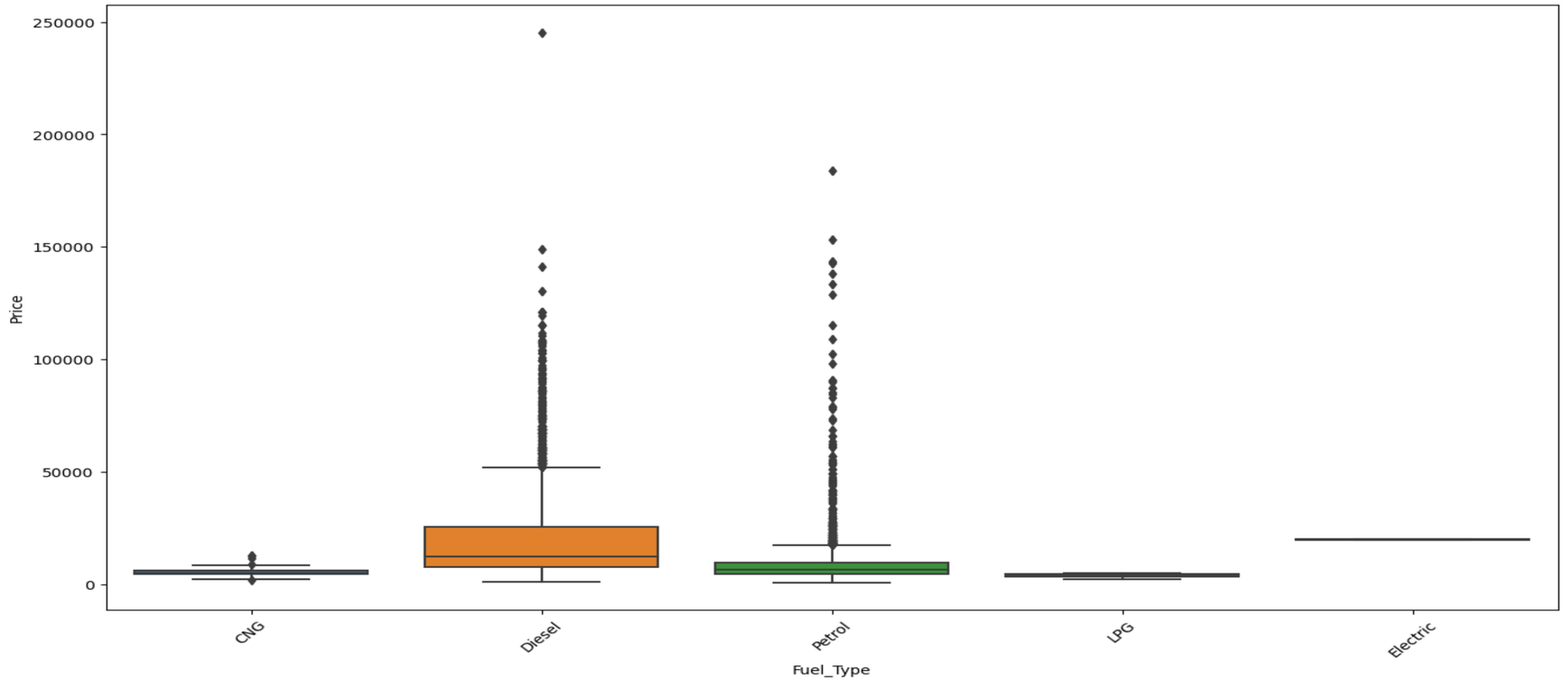
탐색적 분석



목표변수(Price)와 범주형 설명변수(Location)의 분포확인

지역간 가격차이는 거의 없음을 알 수 있다

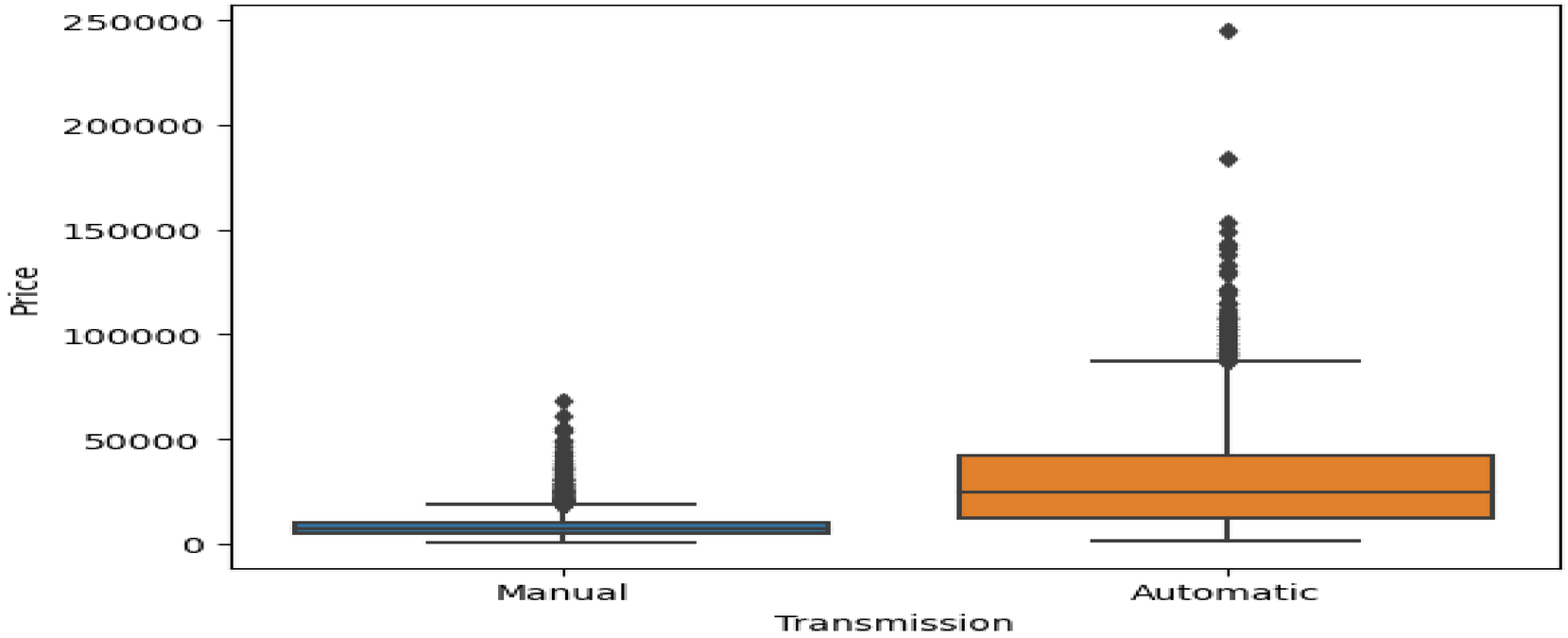
탐색적 분석



목표변수(Price)와 범주형 설명변수(Fuel_Type)의 분포확인

전기차 > 디젤 > petrol > LPG > CNG 순으로 가격대가 형성되어 있다

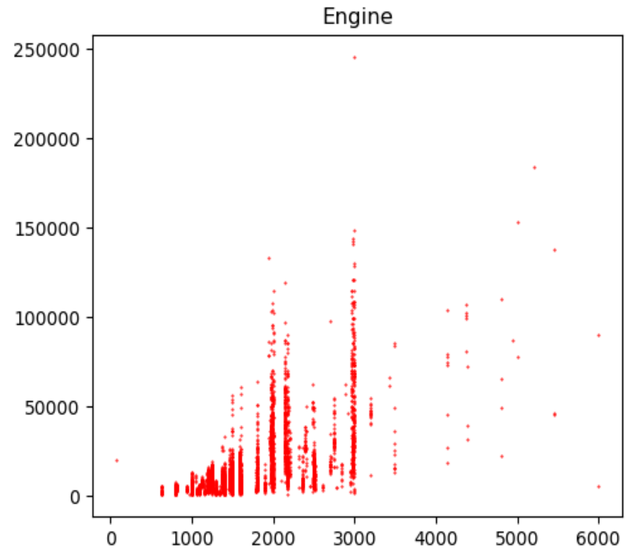
탐색적 분석

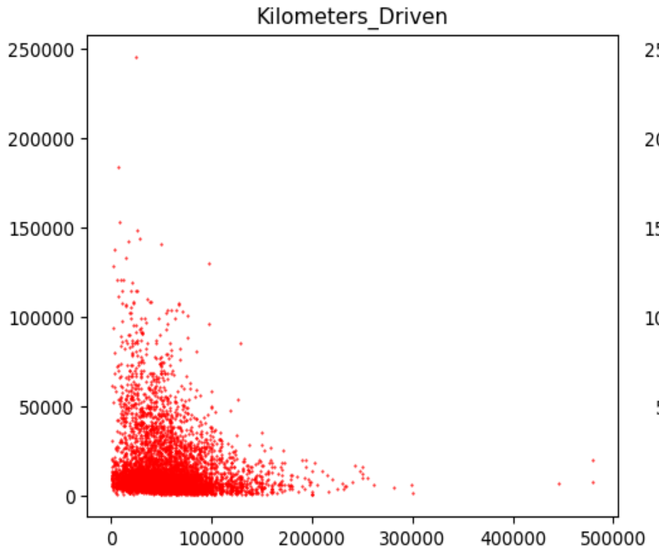


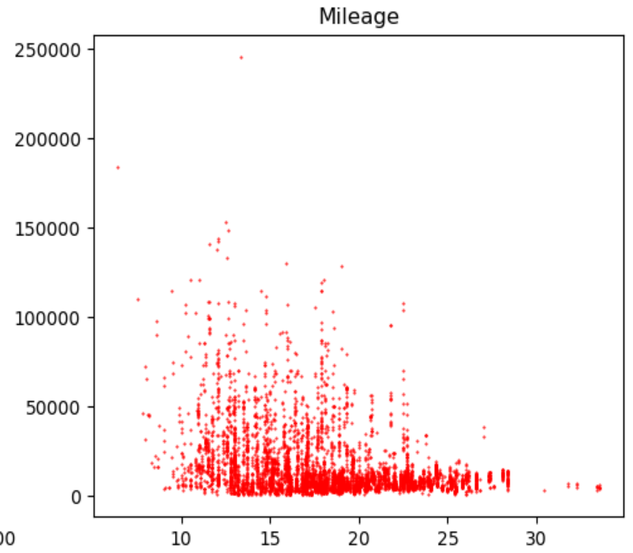
목표변수(Price)와 범주형 설명변수(Transmission)의 분포확인

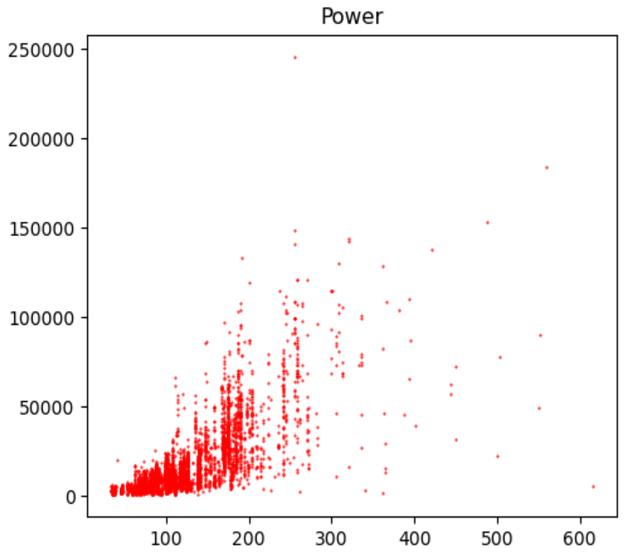
자동 변속기가 수동 변속기보다 비싼 것을 알 수 있다

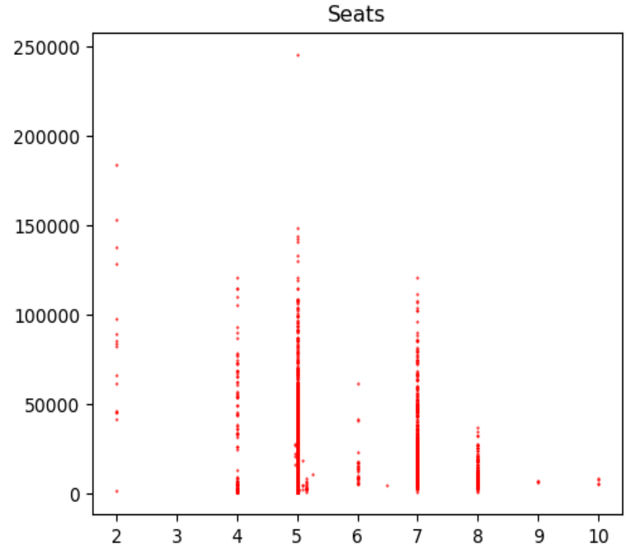
탐색적 분석

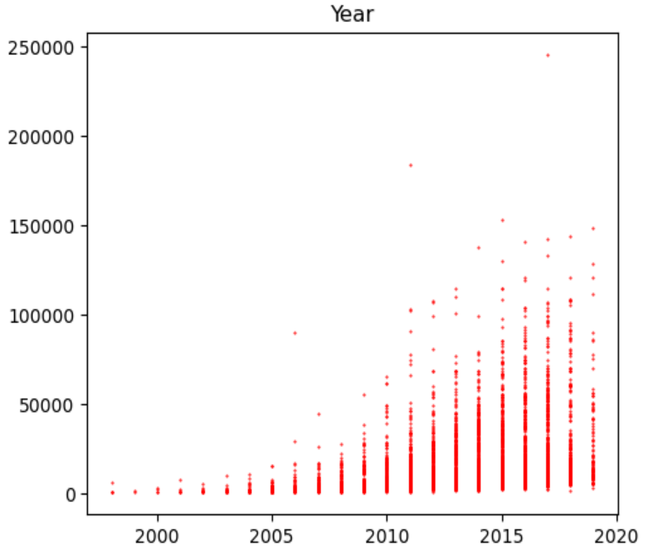








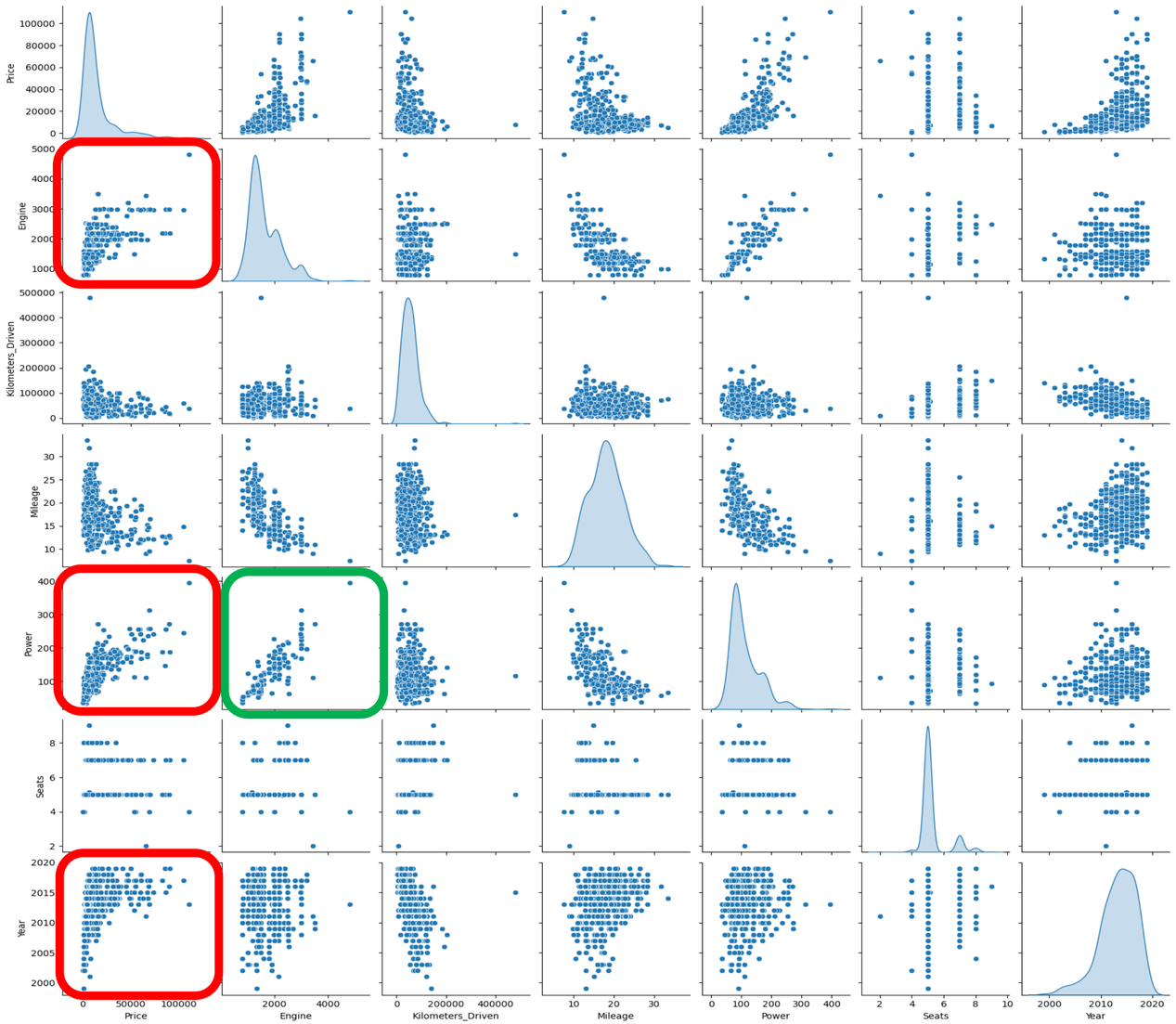




목표변수(Price)와 연속형 설명변수의 분포확인

Mileage는 정규분포가 띄는게 눈에 보인다

탐색적 분석



목표변수(Price)와 연속형 설명변수의 분포확인

현재 목표변수 Price는 설명변수 Engine, Power, year와 상관성이 있음을 알 수 있다

또 Engine과 Power사이에도 상관성이 있음을 보인다

탐색적 분석

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.757
Model:	OLS	Adj. R-squared:	0.755
Method:	Least Squares	F-statistic:	364.3
Date:	Sun, 05 Mar 2023	Prob (F-statistic):	0.00
Time:	19:26:32	Log-Likelihood:	-64235.
No. Observations:	6126	AIC:	1.286e+05
Df Residuals:	6073	BIC:	1.289e+05
Df Model:	52		
Covariance Type:	nonrobust		

Omnibus:	4707.929	Durbin-Watson:	1.980
Prob(Omnibus):	0.000	Jarque-Bera (JB):	445267.096
Skew:	2.998	Prob(JB):	0.00
Kurtosis:	44.334	Cond. No.	5.96e+07

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.579
Model:	OLS	Adj. R-squared:	0.577
Method:	Least Squares	F-statistic:	299.7
Date:	Sun, 05 Mar 2023	Prob (F-statistic):	0.00
Time:	19:26:32	Log-Likelihood:	-65920.
No. Observations:	6126	AIC:	1.319e+05
Df Residuals:	6097	BIC:	1.321e+05
Df Model:	28		
Covariance Type:	nonrobust		

Omnibus:	4163.276	Durbin-Watson:	1.970
Prob(Omnibus):	0.000	Jarque-Bera (JB):	154003.981
Skew:	2.756	Prob(JB):	0.00
Kurtosis:	26.937	Cond. No.	445.

```
chi-square test
chisq: 171528.000
p: 0.000
degree pf freedom:784
```

Brand	Ambassador	Audi	BMW	Bentley	Chevrolet	Datsun	Fiat	Force	Ford	Honda	...	Mini	Mitsubishi	Nissan	Porsche	Renault	Skoda	Tata	T
Ambassador	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Audi	0	239	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
BMW	0	0	272	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Bentley	0	0	0	2	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Chevrolet	0	0	0	0	121	0	0	0	0	0	...	0	0	0	0	0	0	0	
Datsun	0	0	0	0	0	17	0	0	0	0	...	0	0	0	0	0	0	0	
Fiat	0	0	0	0	0	0	30	0	0	0	...	0	0	0	0	0	0	0	
Force	0	0	0	0	0	0	0	3	0	0	...	0	0	0	0	0	0	0	
Ford	0	0	0	0	0	0	0	0	303	0	...	0	0	0	0	0	0	0	
Honda	0	0	0	0	0	0	0	0	0	619	...	0	0	0	0	0	0	0	
Hyundai	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
ISUZU	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Jaguar	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Jeep	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Lamborghini	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Land	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Mahindra	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Maruti	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Mercedes-Benz	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Mini	0	0	0	0	0	0	0	0	0	0	...	29	0	0	0	0	0	0	
Mitsubishi	0	0	0	0	0	0	0	0	0	0	...	0	34	0	0	0	0	0	
Nissan	0	0	0	0	0	0	0	0	0	0	...	0	0	95	0	0	0	0	
Porsche	0	0	0	0	0	0	0	0	0	0	...	0	0	0	18	0	0	0	
Renault	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	151	0	0	
Skoda	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	179	0	
Tata	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	195	
Toyota	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Volkswagen	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
Volvo	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	

설명변수 항목과 목표변수간 차이 검정

회귀분석을 통해 t통계량,
F통계량을 모두 알 수 있다

제조사별 중고차 가격 차이 검정 -> ANOVA 검정

분석 결과 p값이 0보다 작으므로 회귀 모델로서 적합하다

설명력은 57.7%로, **F통계량이 높아 설명력이 좋음을 알 수 있다**

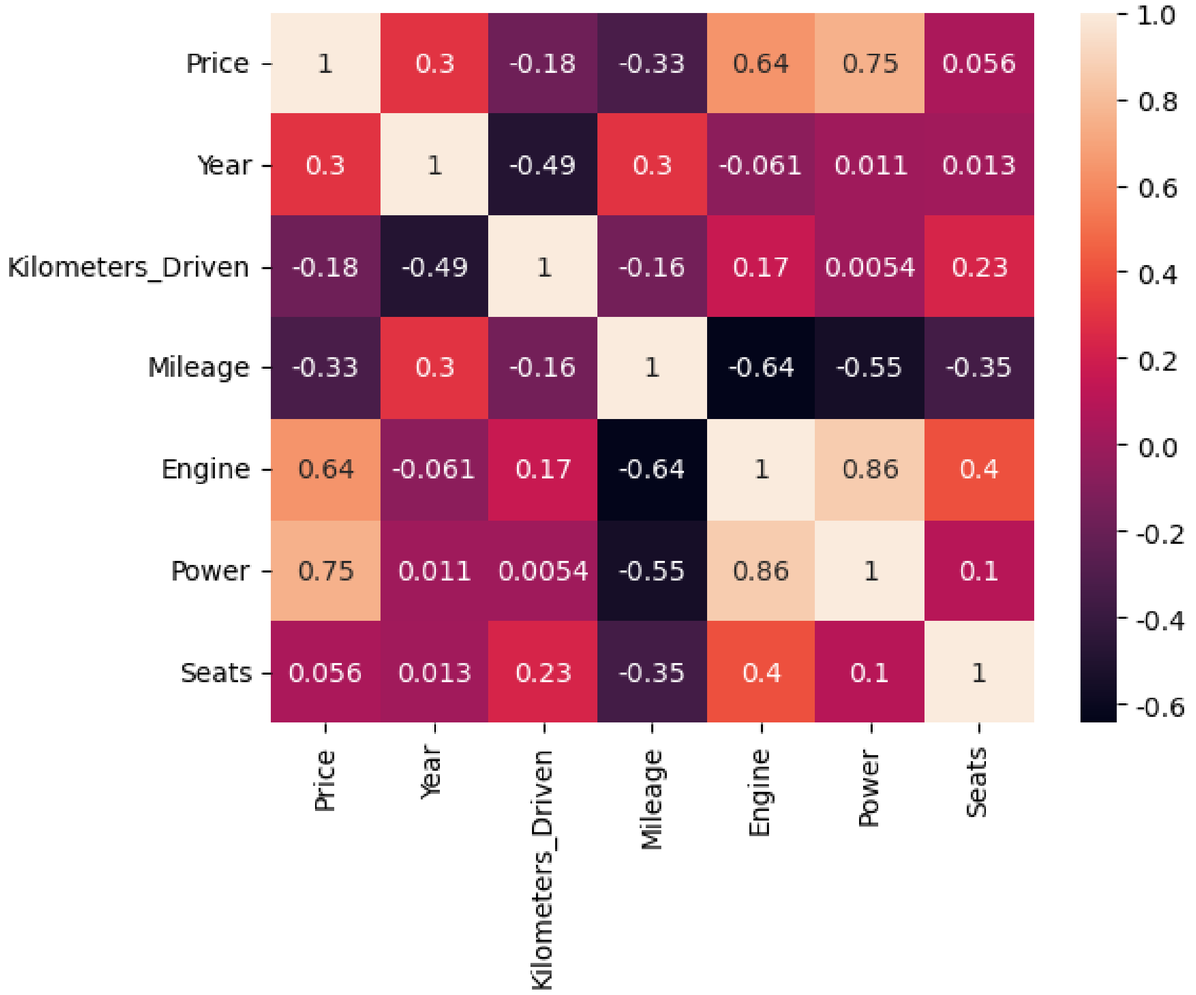
하지만 Prob(Omnibus)와Prob(JB)를 보면 0.05이하로써
정규성이 없음을 판단할 수 있다

따라서 **선형회귀분석은 의미가 없다는 것을 알 수 있다**

제조사별 중고차 비율차이 검정 ->Chi검정

p값이 0.05보다 작으므로, 각
'제조사별로 차량의 갯수 다르다'라는
가설이 설명력이 있음을 알 수 있다

상관 분석



Price는 Engine과 Power와 양의 상관성을 가집니다

즉, 엔진의 배기량과 최대출력이 좋을수록

가격이 높아가는 것을 알 수 있습니다

Mileage는 Engine과 Power와

음의 상관성을 가집니다

즉 엔진의 배기량과 최대출력이 낮을수록

연비가 좋아지는 것을 알 수 있습니다

Power와 Engine은

강한 양의 상관성을 가지고 있습니다

즉, 엔진의 배기량이 클수록

최대출력이 좋음을 알 수 있습니다

변수 선택

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.656
Model:	OLS	Adj. R-squared:	0.656
Method:	Least Squares	F-statistic:	2917.
Date:	Sun, 05 Mar 2023	Prob (F-statistic):	0.00
Time:	19:26:33	Log-Likelihood:	-65303.
No. Observations:	6126	AIC:	1.306e+05
Df Residuals:	6121	BIC:	1.306e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.949e+06	9.43e+04	-31.275	0.000	-3.13e+06	-2.76e+06
Power	224.5699	5.039	44.569	0.000	214.692	234.447
Engine	2.1945	0.459	4.776	0.000	1.294	3.095
Kilometers_Driven	-0.0313	0.005	-6.877	0.000	-0.040	-0.022
Year	1458.7381	46.782	31.182	0.000	1367.029	1550.447

Omnibus:	4045.757	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	311172.601
Skew:	2.397	Prob(JB):	0.00
Kurtosis:	37.585	Cond. No.	4.76e+07

다중회귀분석을 통한 변수선택

p-value가 0.05보다 작아지며,
분산팽창지수가 10이하가되는
변수선택

Power, Engine,
Kilometers_Driven,Year선택

모델링

Grid Searh

```
best estimator model:
DecisionTreeRegressor(min_samples_leaf=10, min_samples_split=10)

best parameter:
{'min_samples_leaf': 10, 'min_samples_split': 10}

best score:
0.781
```

```
best estimator model:
RandomForestRegressor(max_features=2, n_estimators=30)

best parameter:
{'max_features': 2, 'n_estimators': 30}

best score:
0.792
```

```
best estimator model:
GradientBoostingRegressor(learning_rate=0.12, max_depth=10, min_samples_leaf=20)

best parameter:
{'learning_rate': 0.12, 'max_depth': 10, 'min_samples_leaf': 20}

best score:
0.82
```

Decision Tree

최적의 하이퍼 파라미터는

Min_samples_leaf : 10

Min_samples_split : 10

최고 성능 : 78.1 %

RandomForest

최적의 하이퍼 파라미터는

Max_features : 2

n_estimators : 30

최고 성능 : 79.2 %

Gradient Boosting

최적의 하이퍼 파라미터는

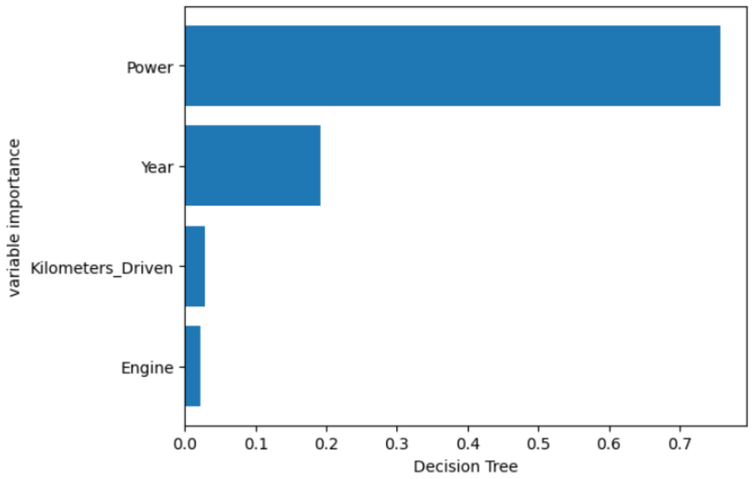
Min_samples_leaf : 20

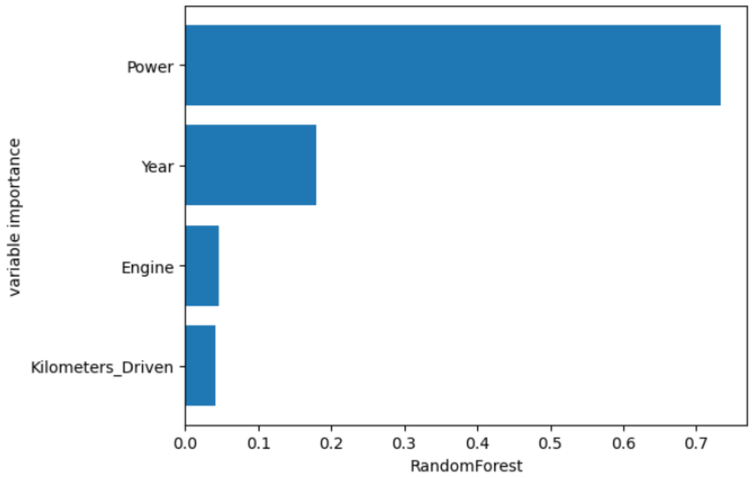
Max_depth : 10

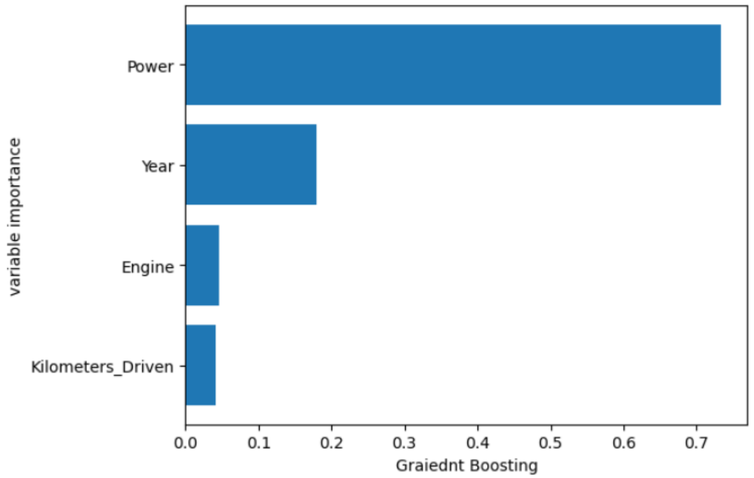
Learning_rate : 0.12

최고 성능 : 82 %

변수 중요도







Decision Tree

Power : 0.757

Year : 0.192

Kilometers_Driven : 0.029

Engine : 0.022

Power > Year > Kilo > Engine

RandomForest

Power : 0.684

Year : 0.180

Kilometers_Driven : 0.083

Engine : 0.053

Power > Year > Kilo > Engine

Gradient Boosting

Power : 0.733

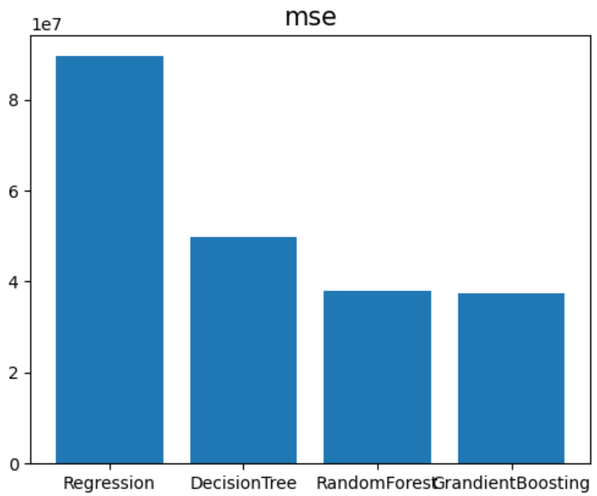
Year : 0.179

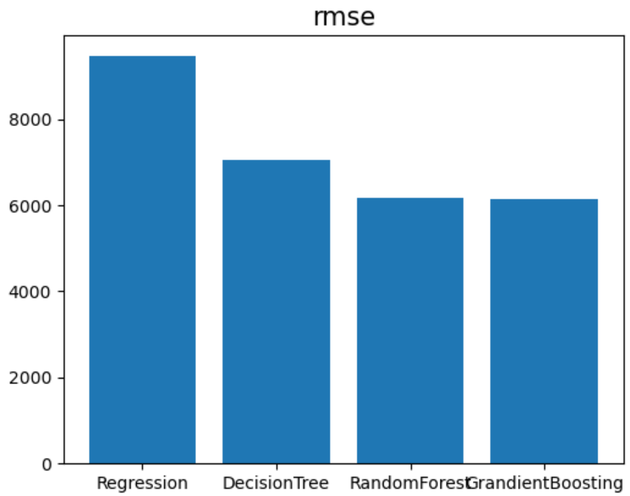
Kilometers_Driven : 0.041

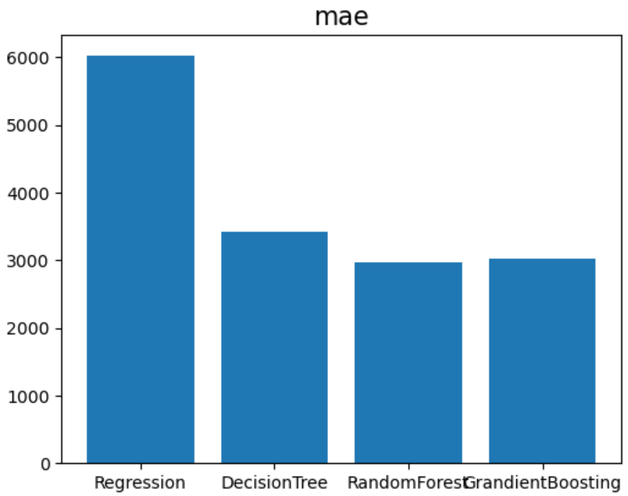
Engine : 0.046

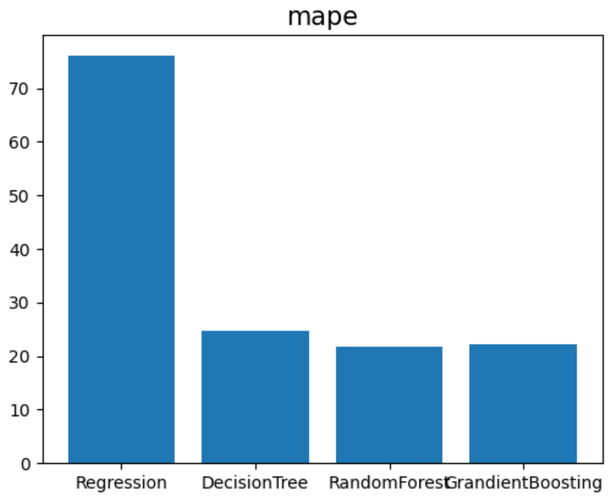
Power > Year > Engine > Kilo

모델평가









회귀분석을 통해 선택된 설명변수를 이용한 모델의 정확도가 가장 낮은 것을 알 수 있다

또한 에러가 가장 낮은 것은 RandomForest와 Gradient Boosting임을 알 수 있다

결론

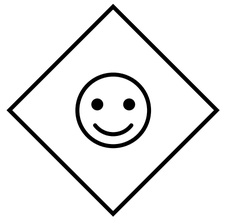
	Regression	DecisionTree	RamdomForest	GradientBoosting
train	0.642	0.859	0.969	0.917
test	0.689	0.827	0.868	0.869

Gradient Boosting 과 RandomForest의 test성능은 비슷하지만, RandomForest는 train/test set의 gap 차이가 많이 나 overfitting 된 것을 볼 수 있다

좋은 성능은 Gradient Boosting > RandomForest > Decision Tree > Regression 순서로 볼 수 있다

종합 결과

변수	변수설명	변수역할	변수형태	분석 제외사유	탐색적 기법			모델링 기법				총점	선정 (사유)
					그래프	검정	상관분석	회귀분석	DT	RF	GB		
Price	중고차 가격 중고차 가격 (단위: 천원)	목표변수	연속형	중고차가격이 100만원 미만 인경우 이상치로 판단 제거 및 결측치 제외									
Name	자동차의 브랜드와 모델	설명변수	연속형	unique한 값이기 때문 (파생변수 Brand생성)									
Location	자동차를 팔거나 구매할 수 있는 위치	설명변수	범주형		blox plot	t							
Year	모델의 년도 혹은 버전	설명변수	연속형		scatter plot scatter matrix	t	o	o	2	2	2	6	선정
Kilometers_ Driven	이전 소유주의 차량 주행거리(Km)	설명변수	연속형	주행거리 50만km이상 이상치로 판단후 제거	scatter plot scatter matrix	t	o	o	3	3	4	11	선정
Fuel_Type	자동차의 사용 연료의 종류	설명변수	범주형		blox plot	t							
Transmission	자동차의 사용 변속기의 종류	설명변수	범주형		blox plot	t							
Owner_Type	소유권이 직접 소유인지, 중고 소유인지 여부	설명변수	범주형		blox plot	t							
Mileage	자동차 회사가 제공하는 표준 주행거리(kmpl)	설명변수	연속형	연비가 0인 경우 이상치로 판단후 제거	scatter plot scatter matrix	t	o						
Engine	엔진의 배기량(cc)	설명변수	연속형		scatter plot scatter matrix	t	o	o	4	4	3	11	선정
Power	엔진의 최대 출력(bhp)	설명변수	연속형		scatter plot scatter matrix	t	o	o	1	1	1	3	선정
Seats	차의 좌석 수	설명변수	연속형	좌석수가 0인 경우 이상치로 판단후 제거	scatter plot scatter matrix	t	o						
New_Price	뉴모델의 가격	설명변수	연속형	영향을 주지않음 (교수님 첨언)									
Brand	자동차의 제조사	설명변수	범주형	파생변수 생성	blox plot	ANOVA, chi							



끝

감사합니다