중고차 가격의 결정적 요소:

회귀분석 팀 프로젝트 결과 보고서

10조

김하은, 이유빈, 장서현, 장윤석, 허재석

목 차

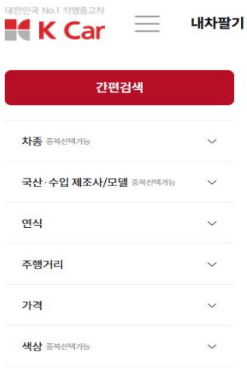
1. 서론

1.1. 데이터 선택 배경

[1]최근 중고차 시장은 기존 개인 딜러 중심에서 '기업형'으로 변화하고 있다. 기존의 케이카, 리본카 등 중고차 플랫폼에 이어 지난해 현대차·기아까지 가세하며 시장의 크기가 커지고 있다. 기업들이 중고차 시장에 뛰어든 이유는 성장 가능성이 충분하기 때문이다. 업계에 따르면 국내 중고차 시장의 거래규모는 신차의 1.4배에 달한다. 한 해에 약 240만대의 중고차가 거래되고 있다.

이처럼 국내 중고차 시장은 크게 성장하며 신차 시장의 거래 규모를 뛰어넘었다. 이러한 시장의 흐름에 따라, 소비자의 관점을 중심으로 중고차 가격에 가장 큰 영향을 끼치는 요인을 분석하고자 한다.

1.2. 중고차 시장 소비자들의 선호 요소



<그림 1> K Car 중고차 검색 화면

중고차 시장에서 제품의 가격은 다양한 변수의 영향을 받아 결정된다. <그림 1>은 실제 중고차 거래 플랫폼 '케이카'에서 중고차를 구매할 때 가격을 결정짓는 변수(필터)이다. 변수로는 차종, 제조사, 연식, 주행거리, 가격, 색상 등이 있다. 이를 통해 실제 중고차 가격을 결정 짓는 요소와 소비자가 고려하는 사항들을 확인할 수 있었고, 이를 참고하여 중고차 가격에 큰 영향을 끼치는 요인에 대해 분석을 진행하였다.

이번 분석은 회귀 분석 과목에서 학습한 회귀 모델의 가정 및 진단 과정을 실제 데이터에 적용하여 중고차 가격 결정 요인을 분석하고, 이를 기반으로 중고차 가격 예측 모델을 만드는 것을 목표로 한다.

---

[1] 김동하. "중고차 시장의 성장 가능성." Smart Financial News, 18 Nov. 2024, https://www.smartfn.co.kr/article/view/sfn202409130005.

2. 데이터 분석

2.1. 데이터 전처리

2.1.1. 분석 데이터셋

"Used Car Price Prediction" 데이터를 분석 데이터셋으로 정했다. 총 6가지의 변수를 가지고 있으며, 설명은 <표 1>과 같다.

변수 이름	타입	설명	표기 예시
Name	String	중고차의 차종	Kia Forte
Year	Integer	차량의 연식	2022
Miles	String	주행 거리	41,406 miles
Color	String	내부와 외부 색상	Gray exterior, Black interior
Condition	String	사고 횟수와 소유주 변경 횟수	No accidents reported, 1 Owner
Price	String	중고차 거래 가격	$15,988

<표 1> 분석 데이터셋 변수 설명

데이터상 결측치는 존재하지 않았으나, Color에 'Unknown'으로 표기되어 있는 것은 결측치로 간주하고 제거하였다. Color는 exterior와 interior로, Condition은 accident reported와 past owners로 분리해 정수형으로 변환했다. 또한 Miles와 Price 값에 포함된 문자열을 제거한 뒤 타입을 Integer로 바꿔주었다. 이와 같이 전처리한 분석 데이터셋은 <표 2>와 같다.

변수 이름	타입	설명	표기 예시
Name	String	중고차의 차종	Kia Forte
Year	Integer	차량의 연식	2022
Miles	Integer	주행 거리	41406
Exterior	String	차량의 외부 색상	Gray
Interior	String	차량의 내부 색상	Black
Past_owners	Integer	차량의 이전 소유주 수	1
Accident_reported	Integer	사고 횟수	0
Price	Integer	중고차 거래 가격	15988

<표 2> 전처리가 완료된 분석 데이터셋 변수 설명

2.1.2. 추가 데이터셋

중고차 가격을 예측하는 모델을 만들기 위해서는 주어진 데이터 외에 추가적인 변수가 필요하다고 판단되었다. 기존 데이터에는 연식, 주행거리, 색상 등의 변수가 포함되어 있었으나, 중고차 가격을 결정짓는 중요한 변수인 신차 가격, 연비 등에 대한 데이터가 부재하였다. 따라서 Car Features and MSRP 데이터셋[2]에서 Make(=Brand) 및 Model 변수를 이용해 차량 제원을 최종 데이터셋에 합치는 것을 고려했다. 추가적인 데이터에서 활용할 변수는 <표 3>과 같다.

변수 이름	타입	설명	표기 예시
Make	String	차량의 제조사	BMW
Model	String	차량의 이름	1 Series
Engine HP	Integer	엔진 마력(엔진의 최대 출력)	335
highway MPG	Integer	고속도로 연비	26
city MPG	Integer	도시 연비	19
MSRP	Integer	권장 소비자 가격	46135

<표 3> MSRP 데이터셋에서 활용할 변수 설명

추가로 확보한 데이터셋에서 같은 차종에 대한 데이터가 여러 개 있는 경우, 차량의 이름(Make와 Model을 합친 이름)을 기준으로 나머지 변수들을 평균 내어 분석 데이터셋에 추가하였다.

2.2. EDA 및 변수 조정

2.2.1. 데이터 선정

위와 같은 데이터에서 변수들 간의 상관성을 고려하고, 범주형 데이터를 적절히 조정하였다.

변수 이름	타입	설명	표기 예시
Name	String	중고차의 차종	Kia Forte
Year	Integer	차량의 연식	2022
Miles	Integer	주행 거리	41406
Exterior	String	차량의 외부 색상	Gray
Interior	String	차량의 내부 색상	Black
Past_owners	Integer	차량의 이전 소유주 수	1
Accident_reported	Integer	사고 횟수	0

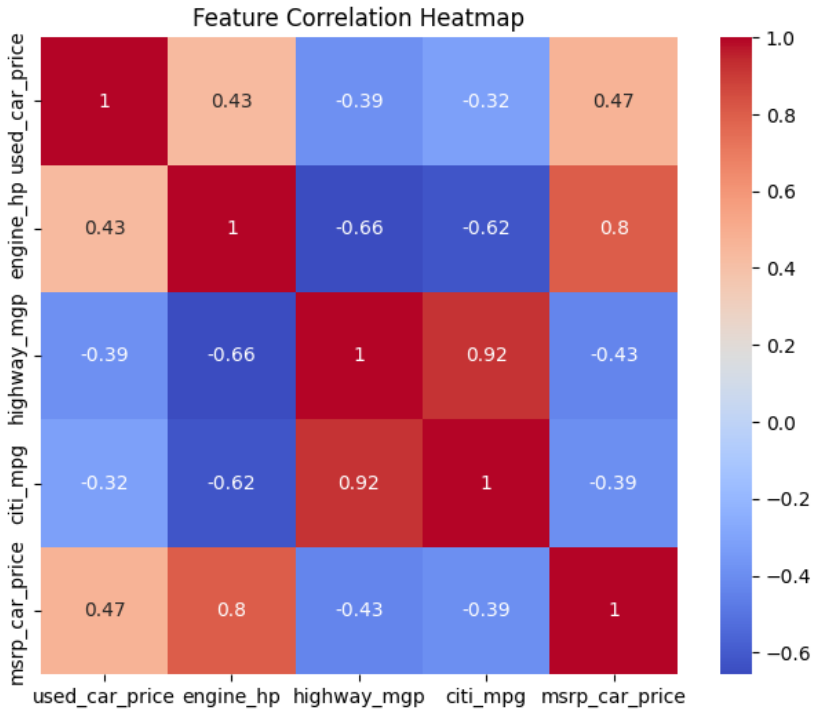
[2] 사용한 데이터셋: Car Features and MSRP. https://www.kaggle.com/datasets/CooperUnion/cardataset

Price	Integer	중고차 거래 가격	15988
Engine HP	Integer	엔진 마력(엔진의 최대 출력)	335
highway MPG	Integer	고속도로 연비	26
city MPG	Integer	도시 연비	19
MSRP	Integer	권장 소비자 가격	46135

<표 4> 전체 분석 데이터셋 변수 설명

2.2.2. 변수 간의 상관관계

○ **MSRP 데이터셋에서 선택한 변수들의 종속변수에 대한 선형성을 확인**



<그림 2> 수치형 변수에 대한 상관계수 행렬

이때 종속변수로 설정한 중고차 가격에 대해 대부분의 예측 변수가 선형관계가 있는 것은 만족스러웠지만, 신차 가격에 대하여 연비나 마력 등의 차량 제원에 관한 데이터들이 선형관계를 보여 이 데이터들을 독립변수로 설정하였을 때, 다중공선성이 발생할 수 있다는 문제점을 발견하게 되었다.

이를 통해 신차가격과 차량 제원에 관한 정보가 상관관계가 있음을 확인할 수 있었고, 추가로 구한 차량 제원 데이터에서는 MSRP(신차가격)만을 선택하여 앞으로의 분석에 활용하기로 결정하였다.

2.2.3. 수치형 데이터

○ **Years와 miles**

일반적으로 차량의 운용 기간이 길수록 주행 거리가 증가하기 때문에, year 변수는 또 다른 독립변수인 주행 거리(miles)와 높은 상관관계를 가질 것으로 예상된다. 또한 우리가 갖고 있는 데이터에는 각 차량의 거래 시점에 대한 정보가 없어 year 변수를 분석에 활용하기 어렵다고 판단하였다.

결과적으로, 차량의 연식을 나타내는 year 변수는 miles 변수와 높은 상관관계를 가질 가능성이 높으며, 데이터의 제약으로 인해 정확한 분석이 어려워 이번 분석의 독립변수에서 제외하였다.

○ **Name과 msrp**

name 변수는 차량의 제조사와 모델명을 포함하고 있어 차량의 브랜드 가치 및 성능에 대한 정보를 내포하고 있다.

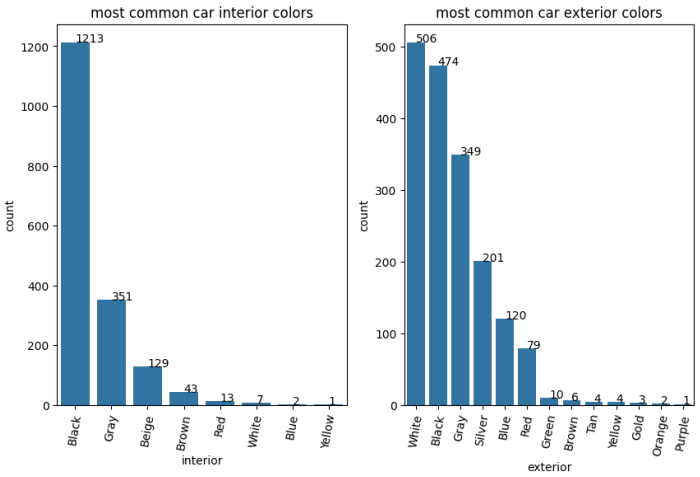
name 변수를 범주형 변수로 변환하여 회귀 분석에 포함시키는 것을 고려했으나, name 변수가 가지는 범주의 수가 지나치게 많아 회귀분석에 적용하기에는 적절하지 않다고 판단했다.

또한, 차량의 제조사와 모델명의 효과는 신차 가격(msrp)에 이미 반영되어 있을 것으로 판단하였다. 즉, name 변수가 msrp 변수와 높은 상관관계를 가질 가능성이 높아 다중공선성 문제를 야기할 수 있다.

이러한 이유로, name 변수는 이번 분석의 독립변수에서 제외하였다.

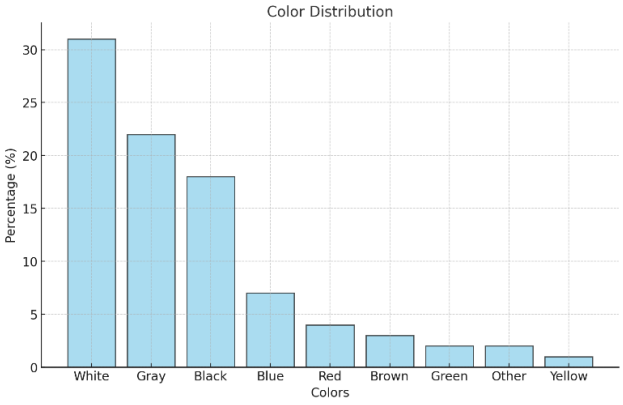
2.2.4. 범주형 데이터

○ **Color (interior, exterior)**



<그림 3> 차량 내외부 색상 분포

주어진 데이터에 결측치(unknown)가 있음을 확인하여 이를 제거하고, 외부 색상에 대해서는 아래의 글로벌 색상 선호도 통계자료를 이용하여 인기색상(White, Gray)과 비인기색상으로 범주화 하였다. 인기색상에 1, 비인기색상에 0을 할당하였다.

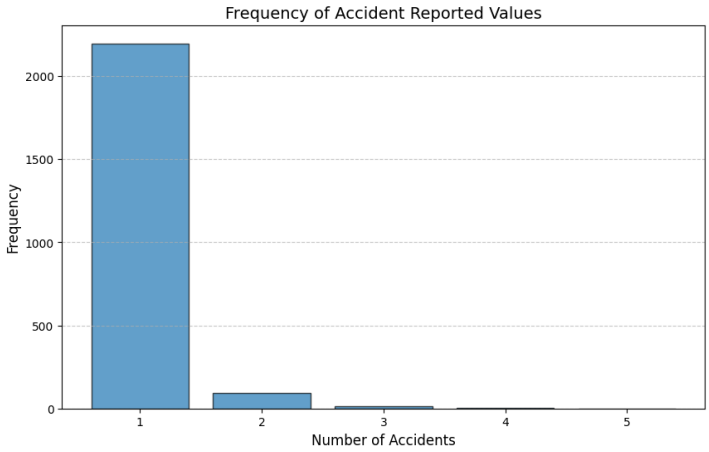


<그림 4> 글로벌 색상 선호도[3]

내부색상에 대해서는 위의 내부 색상 분포에서 많은 비율을 차지하는 상위 2가지 색상(Black, Gray)에 1을 할당하고 나머지 색상에 0을 할당하여 범주화 하였다.

[3] Axalta (자동차 페인트와 관련 재료 및 코팅 분야에서 전문성을 갖춘 미국 기업). 글로벌 자동차 2023 컬러 인기도 리포트. 2023. https://www.axalta.com/kr/ko_KR/ColorCompetency/ColourPopularityReports.html

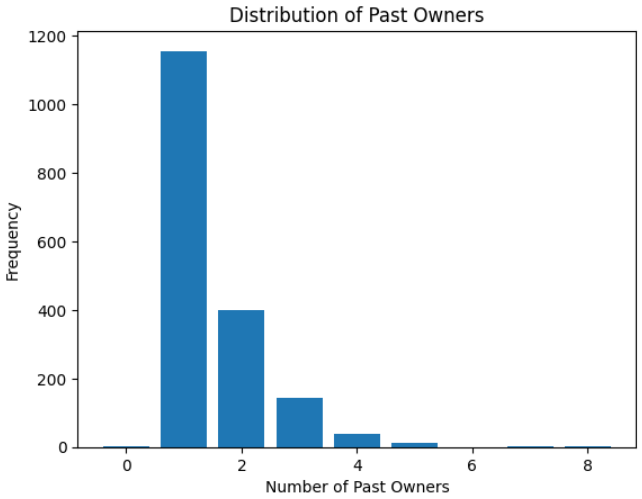
○ **사고 횟수 (accident reported)**



<그림 5> 사고 횟수 분포

주어진 데이터셋에서 각 관측치에 대한 사고 횟수는 1에 많이 분포하고 있었으며 다른 값들은 큰 빈도를 차지하지는 않았다. 따라서 사고 횟수는 '사고가 있었음'과 '사고가 없었음'로 범주화 하여 사고가 있는 관측치에는 1, 사고가 없었던 데이터에는 0을 할당하였다.

○ **이전 소유주 수 (past owners)**



<그림 6> 이전 소유주 수 분포

이전 소유주 수에 대한 데이터는 중고차의 특성 상 이전 소유주 수가 1인 관측치부터 8인 관측치까지 다양하게 분포하고 있다. 이에 2를 기준으로 2 이상인 관측치는 '소유주 수 많음', 2 이하인 관측치는 '소유주 수 적음'로 범주화 하여 각각 1과 0을 할당하였다.

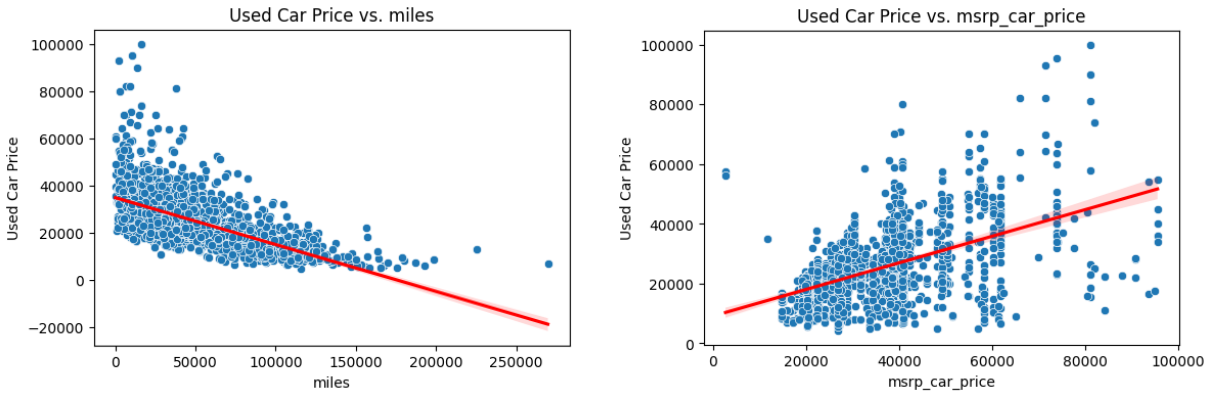
2.3. 최종 변수 선정

다음 표는 앞에서 수행한 전처리 및 변환을 통해 설정한 변수를 정리한 것이다.

변수 종류	이름	타입	설명	표기 예시
설명 변수	miles	Integer	차량 운용 거리	41406
	MSRP car price	Float	신차 권장 소비자 가격	19895.000
	Interior	Integer	내부 색상(빈도)	1
	exterior	Integer	외부 색상(인기 또는 비인기)	1
	Accident reported	Integer	사고 유무	0
	Past owners	Integer	소유주 수 대소	0
종속 변수	Used car price	Integer	중고차 가격	15998

<표 5> 회귀분석에 사용할 변수 정리

2.3.1. 이상치 탐색



<그림 7> 수치형 변수 miles와 msrp car price의 산점도

위 그림으로 최종적으로 선택한 종속변수와 독립변수가 선형성을 가지는지 확인해보았다.

○ **Miles**

중고차 가격과 운용거리의 관계를 보았을 때, 특별한 문제점이 보이지 않았다.

○ **Msrp car price**

중고차 가격과 권장소비자가격의 관계를 산점도로 시각화한 결과, 일부 데이터에서 선형적인 관계를 따르지 않았다. 특히, 중고차 가격이 권장소비자가격보다 높은 경우가 있었으며, 이에 대하여 중고차 시장의 특성을 고려하여 원인을 탐색했다.

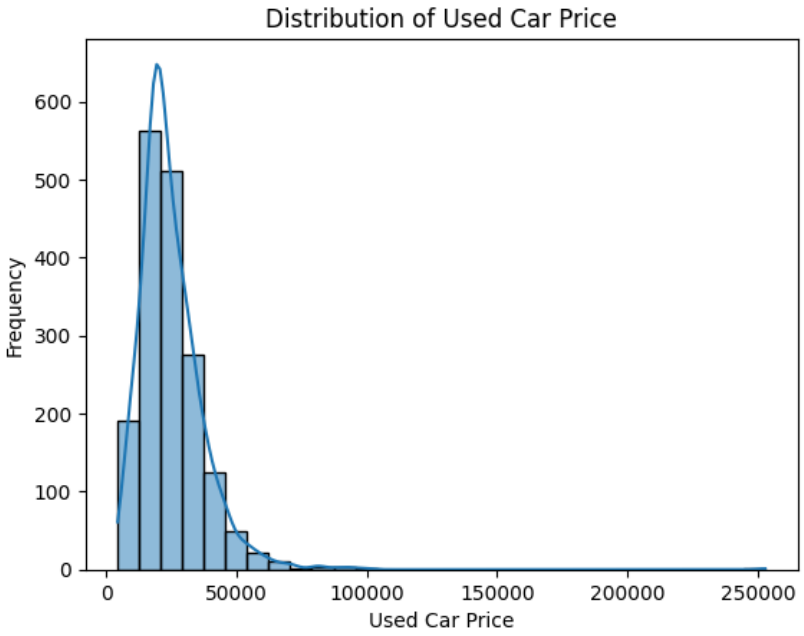
그 중 하나가 차량의 희소성 때문이다. 한정 생산되었거나 단종된 모델의 경우. 수요에 비해 공급이 부족해지면서 중고차 가격이 권장소비자가격을 초과하는 현상이 발생할 수 있다.

분석 결과, 중고차 가격이 권장소비자가격보다 높은 가격으로 거래된 경우 평균적으로 신차가격보다 약 10% 정도의 높은 가격에서 거래되었다. 때문에 이 자료를 토대로 이상치의 기준을 신차 가격의 110%로 보았으며, 이 분석에서는 110% 이상이 되는 중고차 가격은 이상치로 정의하고 제거하였다.

2.3.2. 정규성 확인

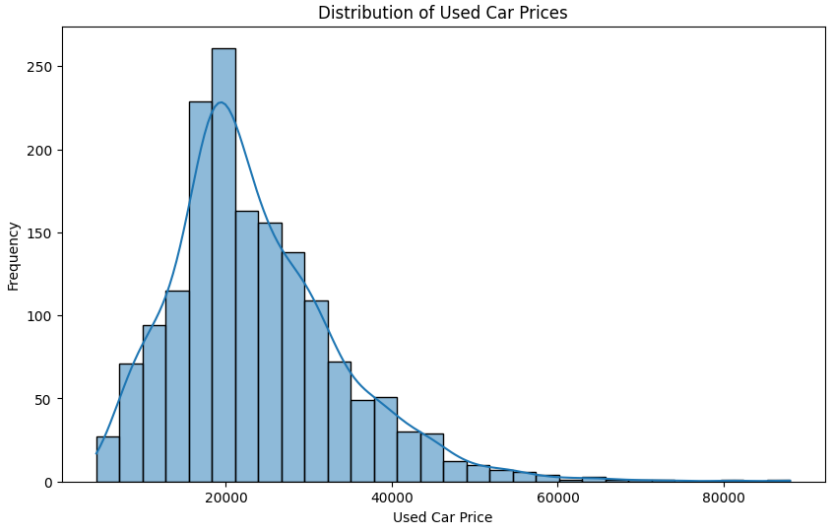
모델의 정확한 추론과 예측 신뢰도를 높이기 위해서 선택한 변수들의 히스토그램을 그려 정규성을 확인해보았다.

○ **Used car price (중고차 가격)**



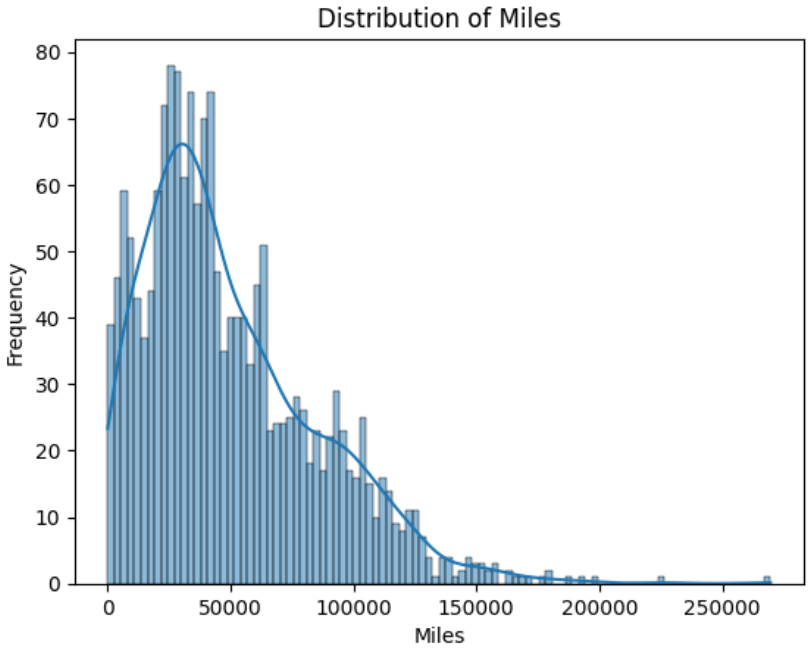
<그림 8> 중고차 가격 히스토그램

종속변수인 used_car_price의 분포가 정규성을 가지지 않음을 확인하였다. 데이터를 살펴보았을 때, 해당 관측치는 롤스로이스의 그랜드 투어러 라인업이었던 롤스로이스 레이스의 컨버터블 버전인 "롤스로이스 던" 모델로, 값이 252900로 확인되었다. 일반적인 중고차 시장의 해석과 예측을 목표로 하는 이번 분석에서는 이러한 관측치들을 적절히 조정하기 위해 중고차 가격에 상한을 100,000으로 설정하여 분석의 정확성을 확보하고자 하였다. 조정 후의 분포는 다음과 같다.



<그림 9> 이상치 조정 후 중고차 가격 히스토그램

○ **Miles (차량 운용 거리)**



<그림 10> 차량 운용 거리 히스토그램

miles 데이터는 대체적으로 정규분포와 비슷한 분포를 보였다. 왼쪽의 값이 거의 없는 것을 확인할 수 있었는데 이는 중고차의 특성 상 miles(운용거리)가 보일 수 있는 특성이라고 생각하여 추가적인 조정은 하지 않았다.

3. 회귀 분석

3.1. 회귀모델 적합

설명변수	계수	표준오차	t-값	p-값	95% 신뢰 구간 하한	95% 신뢰 구간 상한
절편	1.884e+04	700.826	26.887	<.001	1.75e+04	2.02e+04
miles	-0.1486	0.005	-30.215	<.001	-0.158	-0.139
exterior	656.9693	300.642	2.185	0.029	67.285	1246.653
interior	817.8315	486.101	1.682	0.093	-135.615	1771.278
past owners	-2351.2401	360.746	-6.518	<.001	-3058.814	-1643.667
accident reported	-775.6108	364.030	-2.131	0.033	-1489.626	-61.596
msrp car price	0.3493	0.010	35.807	<.001	0.330	0.368

<표 6> OLS 회귀분석 결과

통계량	값	통계량	값	통계량	값
결정 계수(R^2)	0.652	AIC	3.328e+04	관측값 수	1642
수정된 결정 계수	0.651	BIC	3.331e+04	자크-베라 검정	2007.095
F-통계량	510.6	더빈-왓슨 통계량	2.023	자크-베라 p-값	<.001
F-통계량의 p-값	<.001	옴니버스 검정 통계량	250.064	왜도(Skewness)	0.458
Log-Likelihood	-16631.	옴니버스 검정 p-값	<.001	첨도(Kurtosis)	8.338

<표 7> 모형 요약

<표 6>은 앞서 전처리한 데이터로 회귀모델을 적합한 결과, <표 7>은 모형 요약이다. 적합한 회귀모델이 종속 변수(price)의 변동성 중 약 65%를 설명한다. 회귀모델과 관련된 지표의 설명과 설명변수들의 유의성을 검정하기 전에 모델의 각각의 회귀계수에 대한 해석은 다음과 같다.

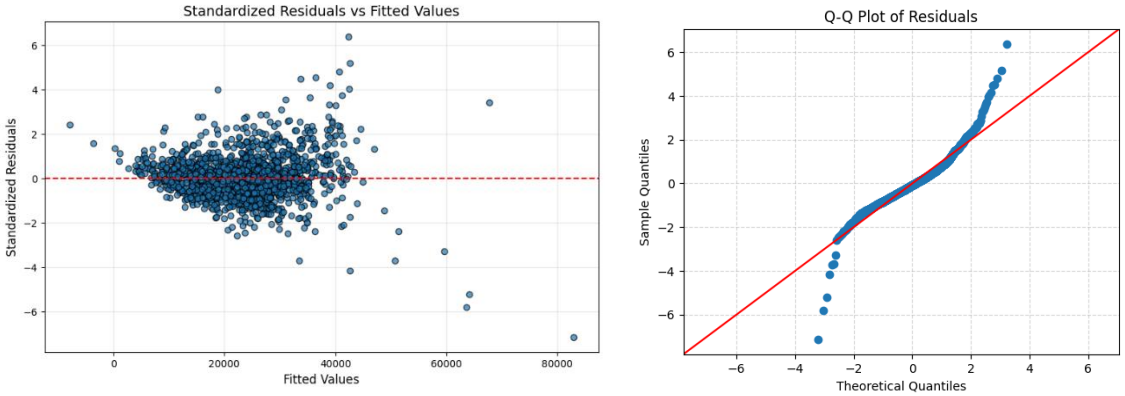
변수	부호	설명
Miles	-	mile이 1 증가할 때, 중고차 가격이 0.15달러씩 감소함을 나타낸다.
Exterior	+	인기 색상을 선택했을 때, 중고차 가격이 약 656달러 더 상승한다.
Interior	+	상위 2개 색상을 선택 시 중고차 가격이 약 817달러 더 상승한다.
Past owners	-	이전 소유주가 많을 경우, 중고차 가격이 약 2351달러 감소한다.
Accident reported	-	사고가 발생했을 때 중고차 가격이 약 775달러 감소한다.
Msrp car price	+	중고차의 가격은 신차가격의 35%정도의 가격에서 다른 변수들의 영향을 받으며 정해진다.

<표 8> 설명변수 각각의 회귀 계수에 대한 해석

이제 각각의 회귀계수의 유의성과 적합된 모델에서 발생할 수 있는 문제를 조정하고자 한다.

3.1.1. 잔차와 관련된 가정 확인 및 특이값 식별

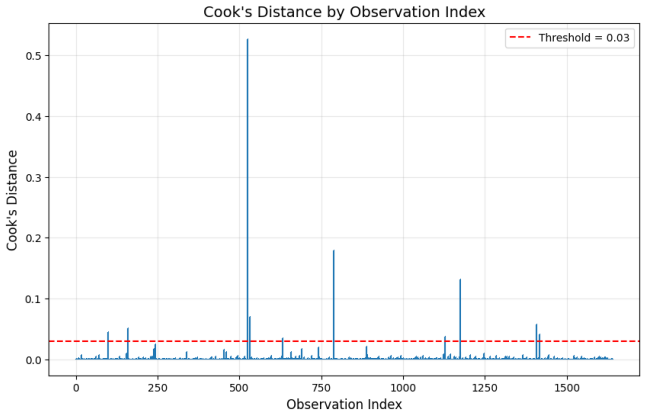
모델이 오차항과 관련된 가정을 위배했는지를 확인하기 위해, 적합된 모델에서 얻은 표준화 잔차의 산점도와 Q-Q Plot을 검토하였다. 이를 통해 오차의 분포와 등분산성을 평가하고자 한다.



<그림 11> 표준화 잔차에 대한 산점도와 Q-Q Plot

표준화 잔차를 시각화한 결과, 잔차들이 독립적이지 않다는 패턴이 관찰되었으며, Q-Q plot에서도 정규성을 보이지 않는다는 것을 확인하였다. 이는 모델의 추가적인 개선이 필요함을 시사한다 이를 해결하기 위해 Cook's Distance로 모델에 큰 영향을 미치는 특이값을 식별하고자 하였다.

3.1.2. 특이값 식별



<그림 12> Cook's Distance 시각화

각 관측치의 쿡의 거리를 계산한 결과, 교재에서 제시한 임계값 1을 기준으로는 특이값을 찾을 수 없었다. 반면, 표본 수를 기반으로 한 일반적인 기준인 4/n을 적용했을 때는 너무 많은 관측치가 특이값으로 분류되어 적절하지 않다고 판단되었다. <그림 12>를 분석한 결과, 상대적으로 큰 값을 가지는 0.03 이상의 관측치를 특이값으로 간주하였다.

3.2. 문제 해결 및 수정된 모델 적합

특이값이 분석의 정확도에 좋지 않은 영향을 준다고 판단하여, 특이값으로 간주된 총 10개의 관측치를 제거하였다. 제거 후의 모델의 적합 결과는 다음과 같다.

설명변수	계수	표준오차	t-값	p-값	95% 신뢰 구간 하한	95% 신뢰 구간 상한
절편	1.709e+04	667.916	25.586	<.001	1.58e+04	1.84e+04
miles	-0.1481	0.005	-32.644	<.001	-0.157	-0.139
exterior	624.3640	276.800	2.256	0.024	81.441	1167.287
interior	1112.2210	451.259	2.485	0.014	227.111	1997.331
past owners	-2285.0073	333.162	-6.859	<.001	-2938.479	-1631.535
accident reported	-678.2604	335.794	-2.020	0.044	-1336.895	-19.626
msrp car price	0.3909	0.010	38.607	<.001	0.371	0.411

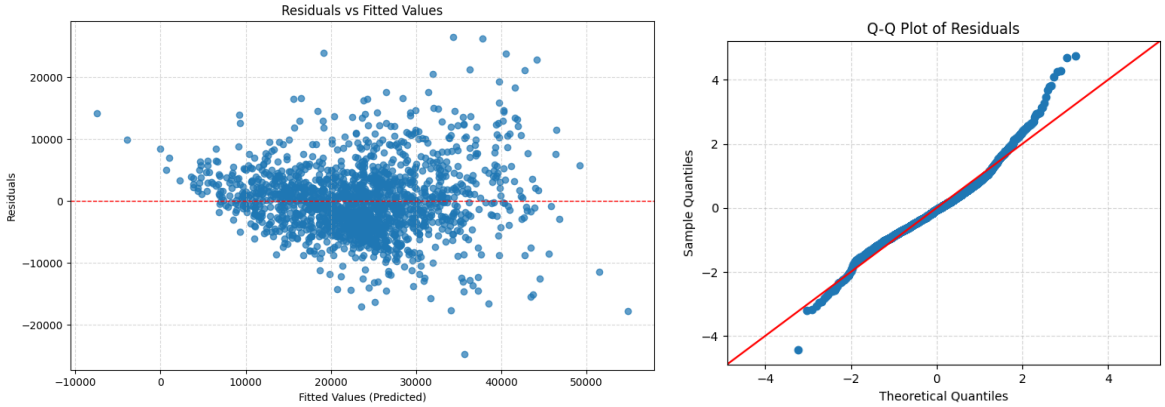
<표 9> 특이값 제거 후 OLS 회귀분석 결과

통계량	값	통계량	값	통계량	값
결정 계수(R^2)	0.685	AIC	3.280e+04	관측값 수	1632
수정된 결정 계수	0.683	BIC	3.283e+04	자크-베라 검정	340.405
F-통계량	587.9	더빈-왓슨 통계량	2.032	자크-베라 p-값	<.001
F-통계량의 p-값	<.001	옴니버스 검정 통계량	147.112	왜도(Skewness)	0.537
Log-Likelihood	-16391.	옴니버스 검정 p-값	<.001	첨도(Kurtosis)	4.963

<표 10> 특이값 제거 후 모형 요약

특이값을 제거한 후 재분석한 결과, 모델의 설명력을 나타내는 지표인 R^2 값이 증가했음을 확인할 수 있었다. 또한, 각 독립변수의 회귀계수가 이전보다 더 유의미한 수준에서 나타나, 모델이 데이터를 더욱 정확히 설명할 수 있게 되었다. 이는 특이값 제거를 통해 모델의 왜곡을 줄이고 예측 및 해석의 신뢰성을 높였음을 의미한다.

특이값을 제거한 모델에서 잔차의 산점도와 Q-Q Plot은 아래 <그림15>와 같으며, 이전과 비교했을 때 잔차의 분포가 더욱 고르게 나타나고 특정 패턴이나 이상치로 인한 왜곡이 줄어든 것을 확인할 수 있다. 이는 특이값 제거가 모델의 잔차가 독립성, 정규성, 그리고 등분산성을 충족하도록 개선했음을 시사하며, 결과적으로 모델의 적합성과 예측 신뢰도가 향상되었음을 보여준다.



<그림 13> 특이값 제거 후 잔차의 산점도와 Q-Q Plot

3.2.1.높은 AIC와 BIC

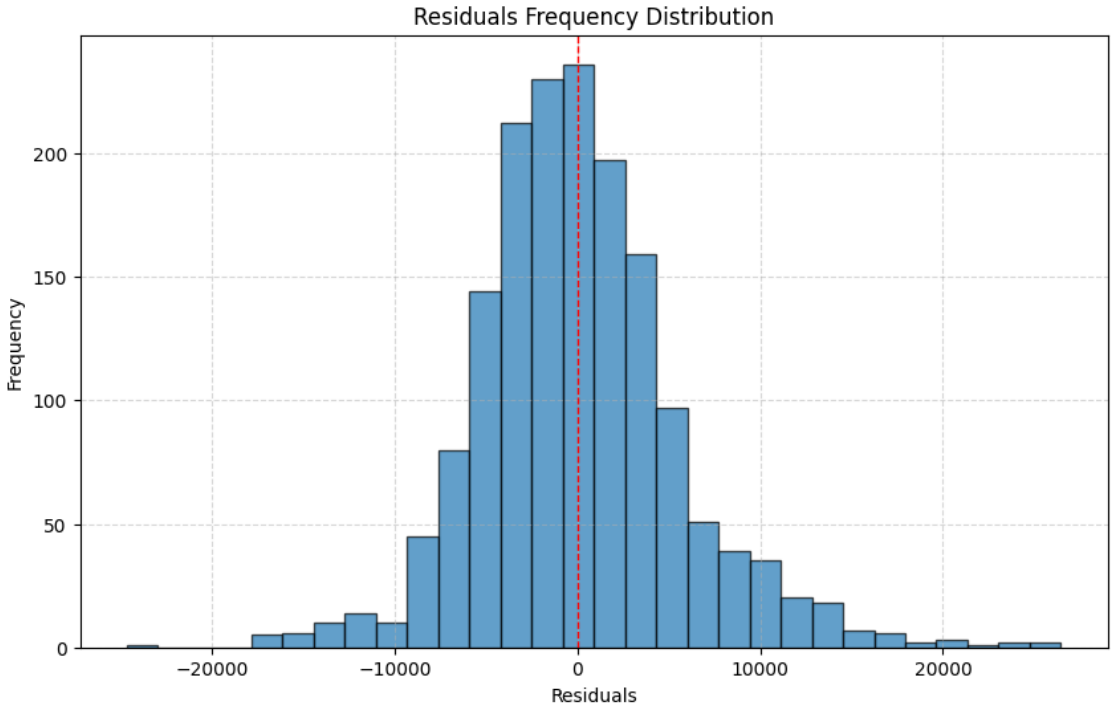
적합된 모델의 결과에서 AIC와 BIC값이 눈에 띄게 큰 것을 확인할 수 있었다. 두 지표는 대표적인 모델의 적합성을 평가하는 기준으로 알려져 있다. 앞서 적합된 모델에서 모델의 유의성을 나타내는 R^2값이 0.68로 준수하고, 각각의 회귀계수도 통계적으로 유의했지만 AIC와 BIC값이 높은 이유에 대해 의문을 가졌다. p-term 방정식이 주어졌을 때, AIC와 BIC의 정의는 다음과 같다.

$$AIC_p = n\,ln(SSE_p/n) + 2p, \qquad BIC_p = n\,ln(SSE_p/n) + p(\ln n)$$

위와 같이 두 지표는 잔차제곱합(SSE), 모델에 포함된 파라미터의 수(p), 데이터의 관측치의 수(n)를 기반으로 계산된다. 앞서 적합한 모델에서 파라미터 수와 관측치 수가 AIC와 BIC값을 과도하게 높이는 요인으로 작용하지 않는다고 판단하였고, 잔차 제곱합(SSE)이 AIC와 BIC를 크게 만드는 주된 원인일 가능성을 탐구하고자 하였다. 잔차 제곱합 SSE는 다음과 같다.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

이 정의를 활용하여 적합된 모델에서 파이썬 코드를 활용해 SSE를 직접 구해보았을 때, 60304357089.40452라는 큰 값이 나왔으며, 이렇게 큰 값이 나온 이유는 종속변수 Y를 중고차 가격으로 설정했기 때문에 실제값과 예측값의 차이가 아래 <그림 16>과 같은 잔차의 히스토그램에서 확인할 수 있듯이 잔차의 값 자체가 약 -20000에서 20000 사이의 값으로 계산됨을 확인할 수 있었다.



<그림 14> 잔차의 히스토그램

위 분석을 통해 AIC와 BIC 값이 크게 나타난 주요 원인은 모델의 잔차 제곱합 (SSE)이 상대적으로 크기 때문이라고 판단하였다. 이는 AIC와 BIC의 정의에서 SSE가 지표 값에 직접적인 영향을 미친다는 점과, SSE가 크면 자연스럽게 모델의 적합도가 낮아 보이도록 계산되는 메커니즘을 고려한 결과다.

이를 검증하기 위해, 종속변수인 중고차 가격을 평균 0, 분산 1로 표준화한 뒤 동일한 독립변수를 사용하여 회귀모델을 다시 적합해 보았다. 표준화를 통해 데이터의 스케일을 조정하여 모델 적합 과정에서 단위 차이에 의한 왜곡을 제거하고자 하였다.

설명변수	계수	표준오차	t-값	p-값	95% 신뢰 구간 하한	95% 신뢰 구간 상한
절편	-0.6416	0.067	-9.526	<.001	-0.774	-0.509
miles	-1.493e-05	4.57e-07	-32.644	<.001	-1.58e-05	-1.4e-05
exterior	0.0630	0.028	2.256	0.024	0.008	0.118
interior	0.1121	0.045	2.465	0.014	0.023	0.021
past owners	-0.2304	0.034	-6.859	<.001	-0.296	-0.165
accident reported	-0.0684	0.034	-2.020	0.044	-0.135	-0.002
msrp car price	3.942e-05	1.02e-06	38.607	<.001	3.74e-05	4.14e-05

<표 11> 종속변수를 표준화한 모형의 OLS 회귀분석 결과

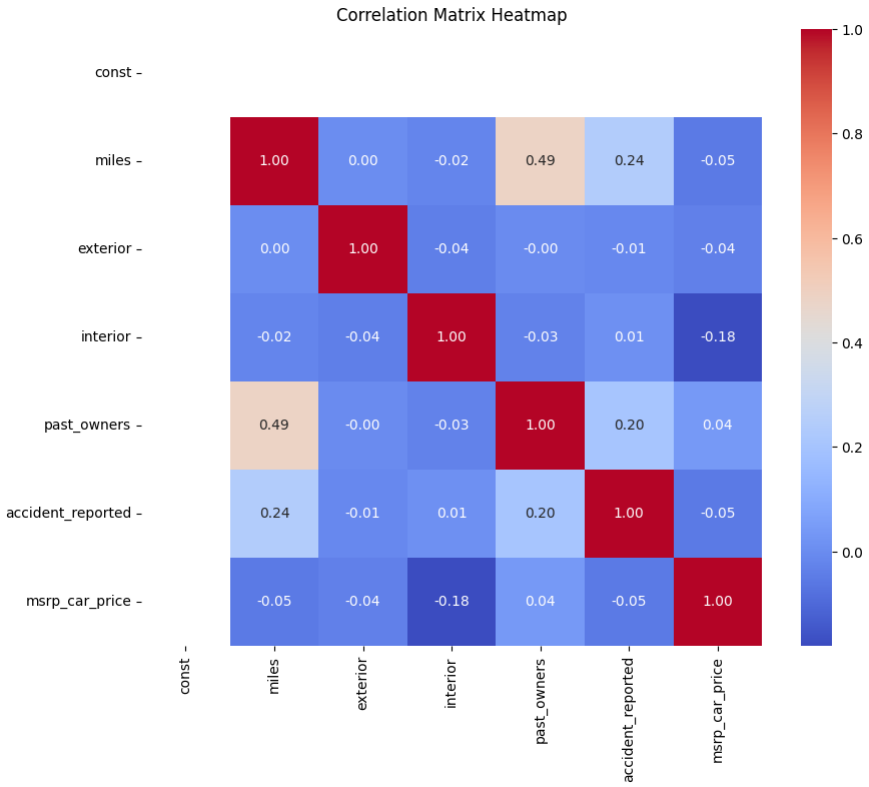
통계량	값	통계량	값	통계량	값
결정 계수(R^2)	0.685	AIC	2761.	관측값 수	1632
수정된 결정 계수	0.683	BIC	2799.	자크-베라 검정	340.405
F-통계량	587.9	더빈-왓슨 통계량	2.032	자크-베라 p-값	<.001
F-통계량의 p-값	<.001	옴니버스 검정 통계량	147.112	왜도(Skewness)	0.537
Log-Likelihood	-1373.5	옴니버스 검정 p-값	<.001	첨도(Kurtosis)	4.963

<표 12> 종속변수를 표준화한 모형 요약

이 결과는 종속변수를 표준화함으로써 데이터의 단위 차이에 의해 발생했던 AIC와 BIC 값의 왜곡이 해소되었음을 보여준다. 표준화 과정으로 잔차의 크기가 조정되었고, 이에 따라 잔차 제곱합(SSE)이 감소하면서 AIC와 BIC 값도 절대적으로 낮아졌다. 이를 통해 AIC와 BIC가 모델의 적합도를 평가하는 지표일 뿐만 아니라, 데이터의 스케일과 잔차의 크기에 민감하게 반응한다는 점을 확인할 수 있었다.

다만, 이러한 변화는 데이터 스케일 조정에 따른 것이며, 모델의 본질적인 설명력이나 적합도의 개선을 의미하지 않는다. 위와 같은 분석은 원래 모델에서 AIC와 BIC 값이 과도하게 높게 나타난 원인을 설명하고, 이러한 지표의 해석 시 데이터 스케일이 영향을 줄 수 있다는 점을 확인하는데 초점을 맞추었다.

3.2.2.변수들의 상관관계 및 VIF



<그림 15> 최종 모델의 상관계수 히트맵

최종 모델의 상관계수 히트맵에서 대부분의 변수들이 낮은 상관계수를 보여, 변수 간 선형적 관계가 강하지 않음을 확인하였다

변수	회귀계수	VIF
Miles	-0.1481	3.91
Exterior	624.364	1.84
Interior	1112.221	4.61
Past owners	-2285.007	2.06
Accident reported	-678.260	1.41
Msrp car price	0.3909	4.24

<표 13> 각 변수의 회귀계수 및 VIF

위 결과에서 모든 변수의 VIF 값이 10을 초과하지 않으며, 사용한 독립 변수들 간에 다중공선성 문제가 없음을 나타낸다. 따라서 회귀분석 모델의 추정치가 신뢰할 수 있으며, 적절한 변수를 선택하였다고 판단할 수 있다.

4. 결론

이번 분석을 통해 중고차의 가격에 영향을 미치는 주요 요인을 식별하고, 이를 기반으로 회귀 모델을 적합하였다. 결과적으로 선정한 데이터셋에서 중고차 가격은 운행거리, 내부 색상, 사고 유무, 이전 소유주 수, 신차 가격과 연관이 있음을 확인하였다.

변수	회귀계수	해석	p-값
Miles	-0.1481	운행 거리 1 증가 시 중고차 가격 약 0.15달러 감소	<.001
Exterior	624.364	인기 색상 선택 시 중고차 가격 약 624달러 상승	0.024
Interior	1112.221	상위 2개 색상 선택 시 중고차 가격 약 1112달러 상승	0.014
Past owners	-2285.007	소유주가 바뀐 적이 있다면 중고차 가격 약 2285달러 감소	<.001
Accident reported	-678.260	사고가 발생한 적이 있다면 중고차 가격 약 678달러 감소	0.044
Msrp car price	0.3909	중고차 가격은 신차 가격의 약 39.1% 수준에서 결정	<.001

<표 14> 최종적으로 적합한 모델의 해석

이상치 제거를 통해 모델의 안정성을 높였으며, 이는 모델 설명력의 개선과 회귀계수의 통계적 유의성에서 확인되었다. 그러나 데이터의 크기를 더욱 확대하고 다양한 조건에서 중고차 가격 패턴을 분석한다면, 모델의 일반화 가능성을 높이고 더 폭넓은 해석을 제공할 수 있을 것이다.

특히, 중고차 구매에 영향을 미칠 수 있는 추가적인 외부 요인을 데이터에 포함한다면, 예측 모델의 정교함을 더욱 높일 수 있을 것이다. 예를 들어, 소비자 선호도, 지역별 가격 차이, 브랜드 신뢰도, 또는 특정 차량 모델의 희소성 같은 요인을 고려하면 중고차 가격의 복잡한 결정 요인을 더 잘 반영할 수 있다.

마지막으로, 종속변수인 중고차 가격의 단위 스케일이 AIC와 BIC 값에 큰 영향을 미친다는 점을 확인하였다. 이를 해결하기 위해 중고차 가격을 표준화한 뒤 모델을 적합한 결과, AIC와 BIC 값이 낮아지는 것을 확인하였다.

이러한 결과는 AIC와 BIC가 데이터의 스케일과 잔차 크기에 민감하게 반응한다는 점을 보여준다. 다만, 표준화 자체가 모델의 본질적인 설명력이나 적합도를 개선하지는 않으며, AIC와 BIC의 변화는 스케일 조정의 결과임을 유념해야 한다.

이번 분석은 중고차 가격 결정 요인에 대한 통찰을 제공하며, 이를 기반으로 향후 중고차 시장에서 더 나은 예측 모델과 분석 프레임워크를 구축하는 데 기여할 수 있을 것이다.