중고차 판매가 예측 및 최적의 판매사이트 추천

기계학습의 이해 2조

이승우 서준형
심현석 이동욱

목차

01 분석 배경

중고차 거래시장의 성장

2021년 기준, 중고차 거래량은 이미 신차 등록 수의 2배를 추월하였고
중고차 온라인 거래 시장의 규모도 꾸준히 확대

기준을 알 수 없는 중고차 판매가

여전히 중고차 판매가가 어떻게 결정되는지는 명확히 알려져 있지 않으며
거래자들은 중고차 딜러가 제시하는 가격을 믿고 거래에 응할 수밖에 없음

01 분석 주제와 목적

주제	중고차 판매가 예측 및 최적의 중고차 판매사이트 추천
목적	국산 중고차의 차량 정보가 주어졌을 때, 어떤 중고차 거래 사이트에 판매하면 가장 이득을 볼 수 있을지 예측
기대 효과	예측 모델을 통해 중고차 판매자가 어떤 사이트에 판매해야 유리한지 판단 판매가 얼마 정도일지 예측 가능

01 분석 사이트 및 차종 선정

사이트명	전체 등록대수
엔카	163,204
KB차차차	143,046
보배드림	49,476
K Car	9,049
현대글로비스	1,227

▶ 엔카, KB차차차, 보배드림 선정

(2024.03.20. 기준)

분석 차종을 국산차로 한정한 이유

1. 외제차는 일정 기간 임대(리스)하는 형태로 이용하는 경우가 더 많기 때문에 '판매가'를 예측하는 분석 목적에 적절하지 않음
2. 보배드림에 등록된 3,280대의 외제차 중 단 410개의 차량만 가격이 기재되어 있으며, 대다수 매물의 판매가는 상담 후 확정되므로 결측치가 상당수 존재

02 데이터 소개

보배드림 중고차 데이터 (1,732 rows × 22 columns)

예측변수 (수치형)

가격

설명변수 (범주형, 18개)

브랜드	차종	연식	색상	변속기
연료	선루프	LED헤드램프	어댑티드헤드램프	가죽시트
열선시트(앞좌석)	통풍시트	후방센서	스마트키	네비게이션(순정)
네비게이션(비순정)	전손유무	침수유무		

설명변수 (수치형, 3개)

배기량	주행거리	소유자이전횟수

*전손: 자동차가 완전히 파손되어 수리가 거의 불가능한 상태

03 데이터 전처리 전후 비교

Before (1,913 × 21)

	차종	가격	연식	배기량	주행거리	색상	변속기	연료	선루프	LED 헤드램프	어댑티드헤드램프	가죽시트	열선시트.앞좌석.	통풍시트	후방센서	스마트키	네비게이션.순정.	네비게이션.비순정.	전손유무	소유자.이전.횟수	침수유무
1	쉐보레 임팔라 2.5 LTZ	599만원	2016.01	2,457 cc (199마력)	182,427 km	은색	자동	가솔린	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
2	쉐보레 올 뉴 말리부 1.5 터보 LTZ	1,190만원	2018.09	1,490 cc (166마력)	86,000 km	흰색	자동	가솔린	0	0	0	1	1	1	1	1	0	0			
3	제네시스 더 올 뉴 G80 2.5 터보 AWD	3,540만원	2021.03	2,497 cc (304마력)	122,897 km	검정색	자동	가솔린	1	1	1	1	1	1	1	1	1	0			
4	GM대우 알페온 EL240 프리미엄	259만원	2012.08	2,384 cc (181마력)	150,000 km	은색	자동	가솔린	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
5	기아 더 K9 3.8 AWD 플래티넘II	3,450만원	2019.04	3,778 cc (315마력)	48,830 km	흰색	자동	가솔린	1	1	1	1	1	1	1	1	1	0			
6	기아 더 K9 3.8 플래티넘	2,500만원	2019.01	3,778 cc (315마력)	105,000 km	검정색	자동	가솔린	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	전손: 0	1회	침수분손 : 0

After (1,732 × 22)

	브랜드	차종	가격	연식	배기량	주행거리	색상	변속기	연료	선루프	LED 헤드램프	어댑티드헤드램프	가죽시트	열선시트.앞좌석.	통풍시트	후방센서	스마트키	네비게이션.순정.	네비게이션.비순정.	전손유무	소유자.이전.횟수	침수유무
1	쉐보레	임팔라 2.5 LTZ	599	2016.01	2457	182427	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	쉐보레	올 뉴 말리부 1.5 터보 LTZ	1190	2018.09	1490	86000	2	1	1	0	0	0	1	1	1	1	1	0	0	0	0	0
3	제네시스	더 올 뉴 G80 2.5 터보 AWD	3540	2021.03	2497	122897	3	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
4	GM대우	알페온 EL240 프리미엄	259	2012.08	2384	150000	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
5	기아	더 K9 3.8 AWD 플래티넘II	3450	2019.04	3778	48830	2	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
6	기아	더 K9 3.8 플래티넘	2500	2019.01	3778	105000	3	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0

03 예측변수와 설명변수의 전처리

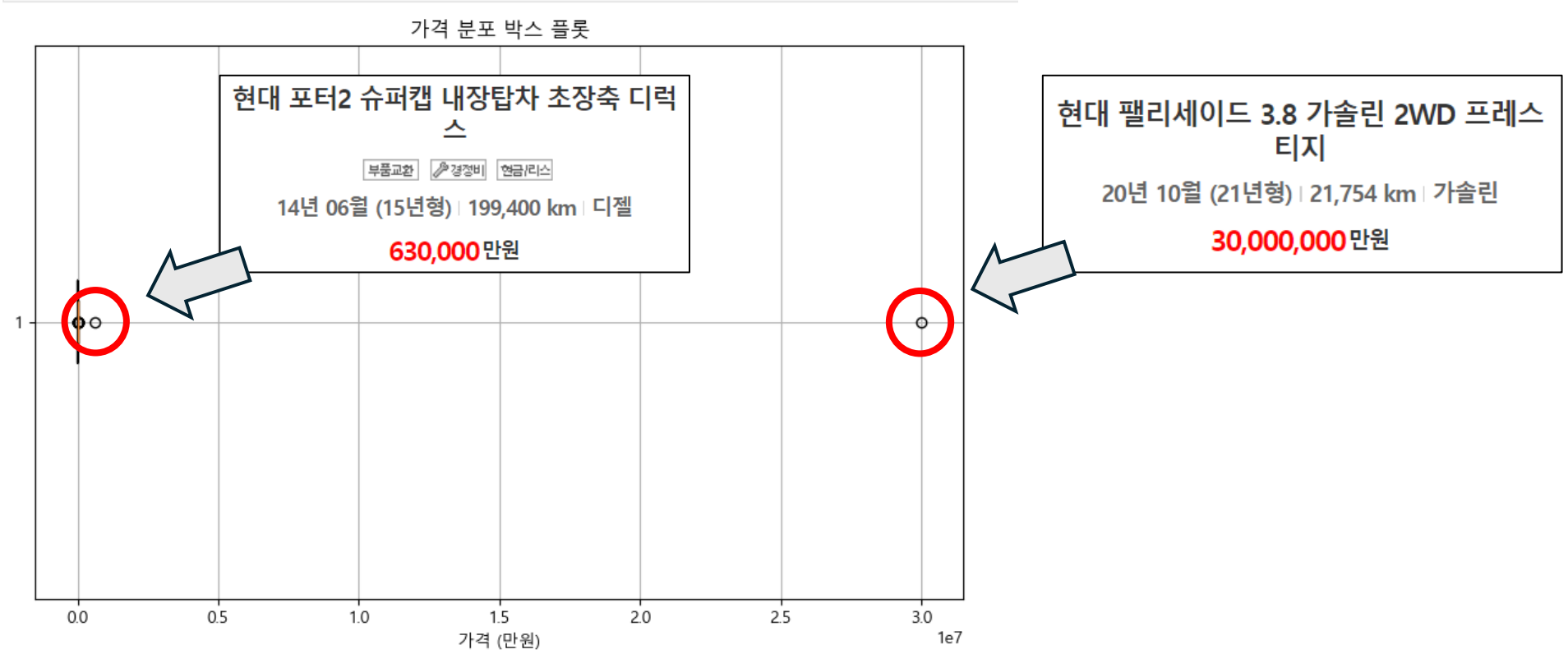
예측변수 데이터의 전처리 절차

- 결측치/텍스트('판매완료' 등) 존재 행 제거
- 가격은 숫자형 데이터로 변환

설명변수 데이터의 전처리 절차

- 차종 열을 첫 공백을 기준으로 브랜드와 차종 열로 분할
- 연식의 서식을 YYYY.MM으로 변환
- 배기량, 주행거리 숫자형 데이터로 변환
- 유사한 색끼리 하나로 묶어준 후 색상, 변속기, 연료에 레이블 인코딩
- 차량 옵션 변수 중 결측치 0으로 대체

03 비정상적인 가격 확인

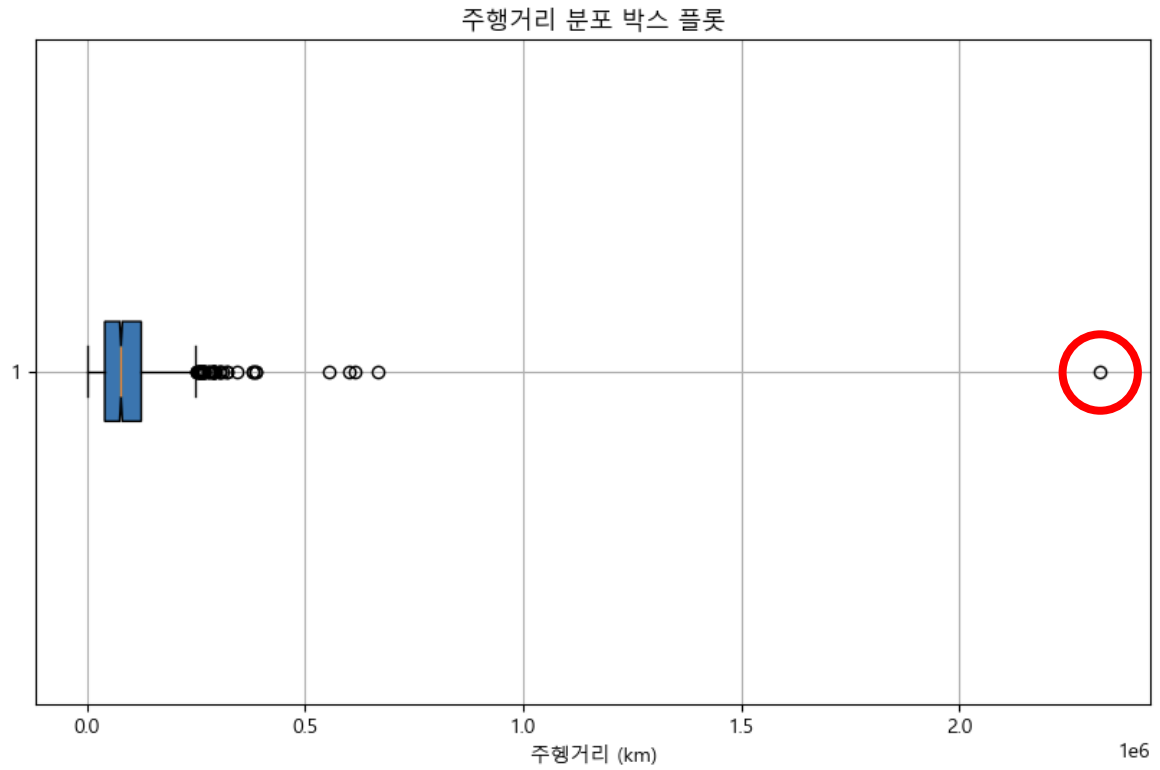


가격 변수 시각화 결과 비정상적인 이상치가 확인

보배드림에서 해당 데이터를 확인한 결과 중고차 판매자가 잘못 기록

30,000,000만원 → 3,000만원으로, 630,000만원 → 630만원으로 데이터 수정

03 비정상적인 주행거리 확인



기본정보

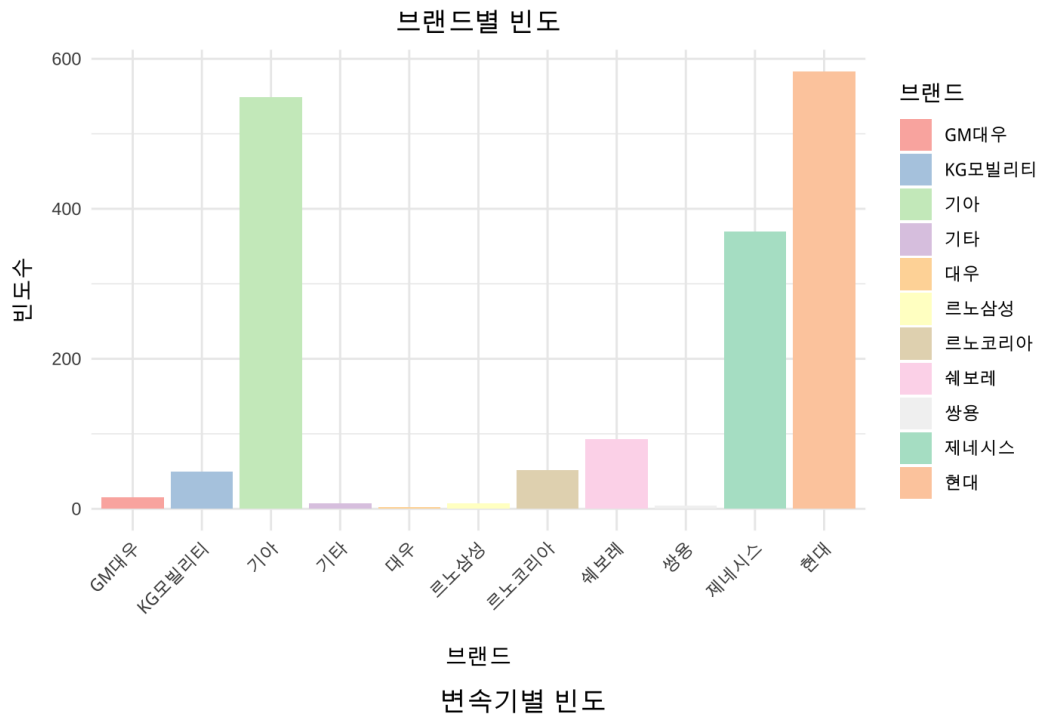
연식	2009.10 (10년형)
주행거리	2,322,500 km

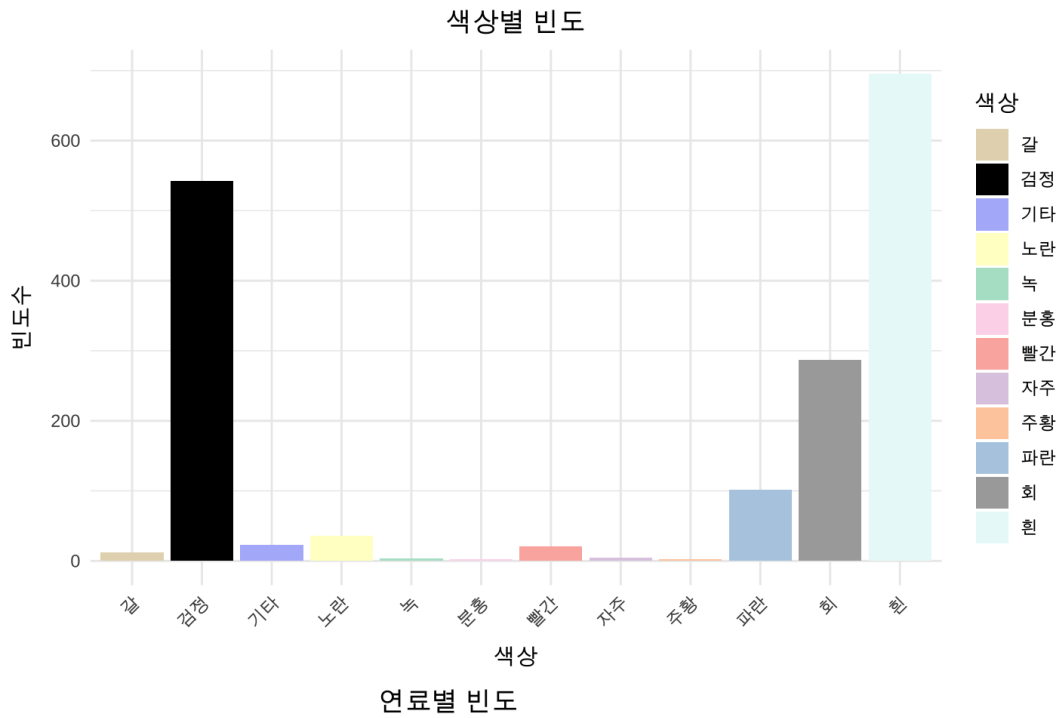
주행거리 변수에 대해 boxplot을 그려본 결과 비정상적인 이상치가 확인

보배드림에서 해당 데이터를 확인한 결과 중고차 판매자가 잘못 기록한 것으로 파악

2,322,500km → 232,000km으로 데이터 수정

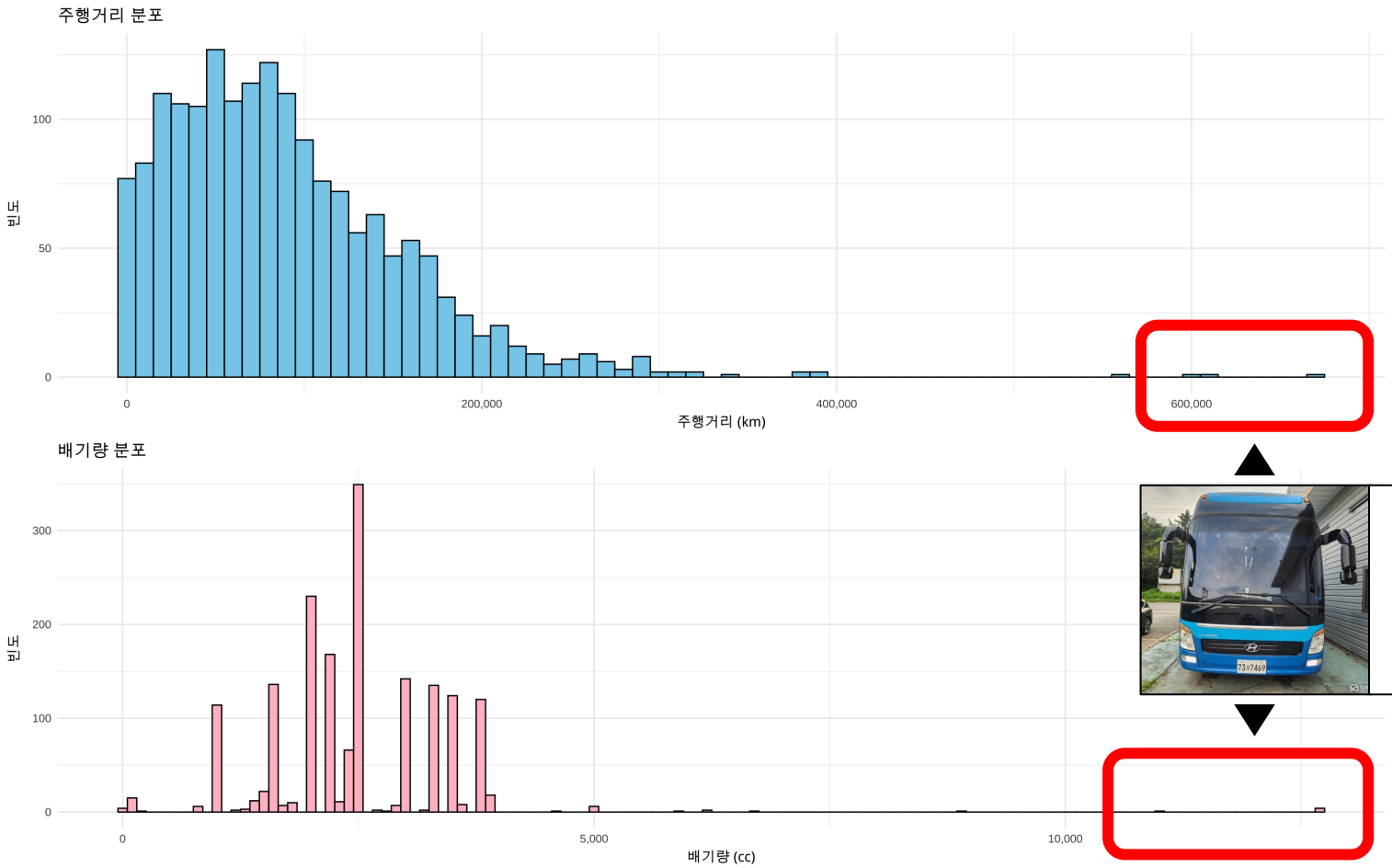
04 브랜드, 색상별 빈도

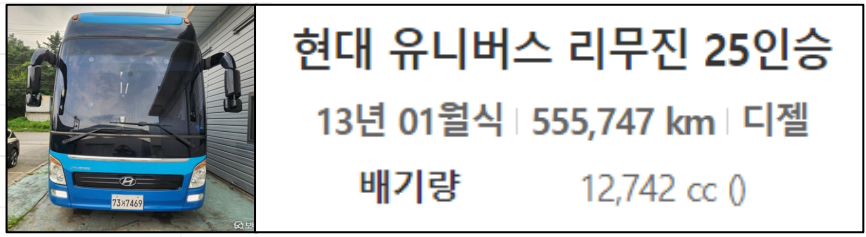




- 브랜드 - 상위 3개 데이터(현대,기아,제네시스)가 전체의 약 86.7%
- 색상 - 상위 3개 데이터(흰색,검정색,회색)가 전체의 약 88.0%
- 데이터 내에 특정 브랜드, 특정 색상의 차량들이 집중

04 주행거리, 배기량의 분포





- 주행거리와 배기량의 분포 모두 오른쪽으로 꼬리가 긴 형태
- 25인승 이상 대형 차량 매물들의 주행거리와 배기량이 모두 높아 분포에 영향

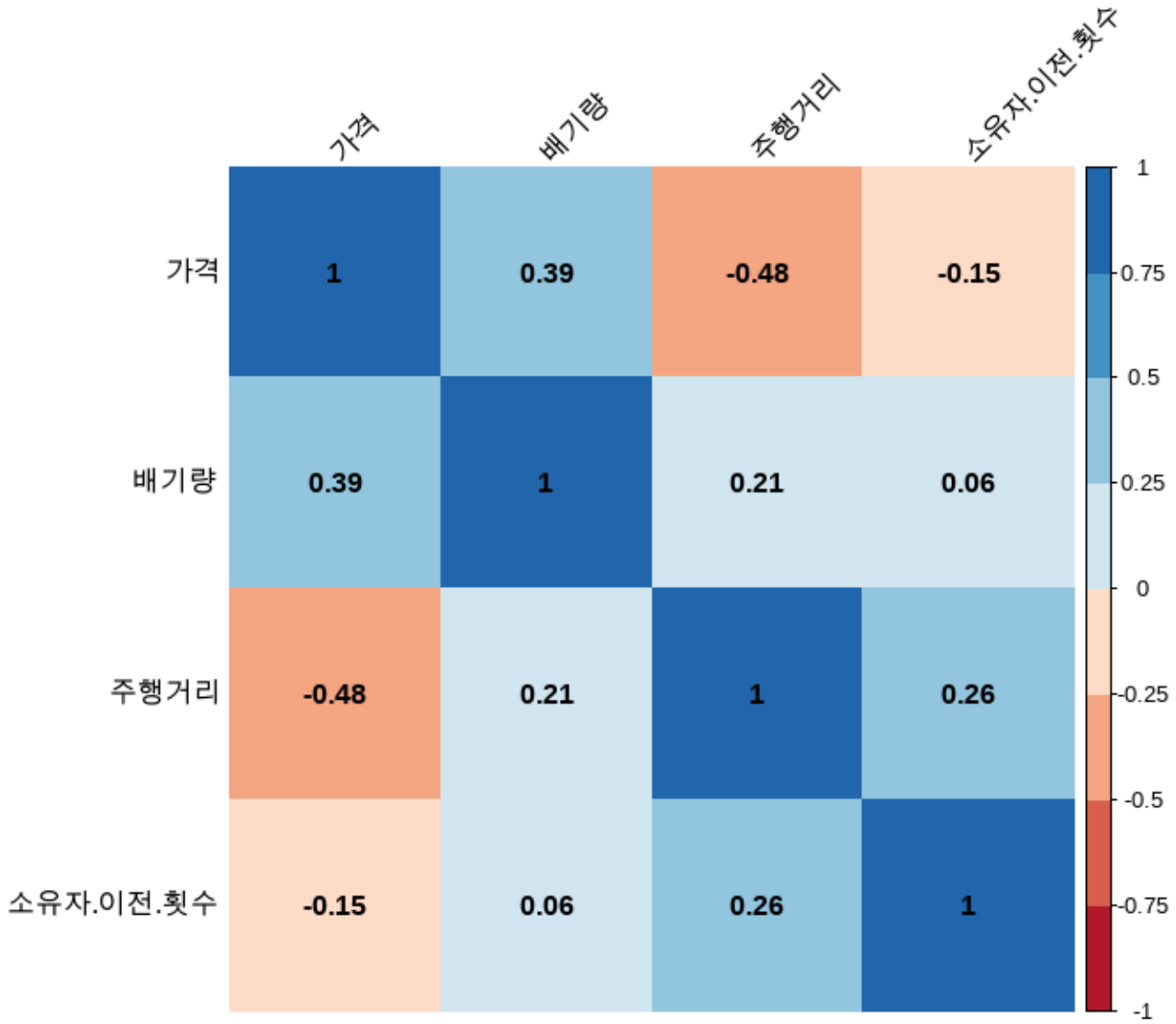
05 예측변수와 범주형 변수 간 관계

- **분석 목적**: 범주형 변수의 수준에 따라 예측변수에 유의한 차이가 있는지 확인
- **분석 방법** (유의수준 = 0.05)
  1. 가격은 정규성 가정을 만족하지 않으므로 일원배치 분산분석과 T-검정을 수행할 수 없음
  2. 비모수 검정법인 Kruskal-Wallis 검정, 맨-휘트니 U 검정을 이용하며 p-value를 계산함
- **귀무가설**: 각 수준에 따른 예측변수 분포의 중앙값은 모두 같다.
- **대립가설**: 각 수준에 따른 예측변수 분포의 중앙값이 모두 같은 것은 아니다.

p-value가 유의수준 0.05보다 작은 변수	p-value가 유의수준 0.05보다 큰 변수
브랜드, 색상, 연료, 변속기, LED헤드램프, 어댑티드헤드램프, 가죽시트, 통풍시트, 네비게이션(순정), 네비게이션(비순정)	선루프, 열선시트(앞좌석), 후방센서, 스마트키

- 유의수준 0.05에서 선루프, 열선시트, 후방센서, 스마트키는 종류에 따라 가격에 유의한 차이가 없다

05 예측변수와 수치형 변수 간 관계



양의 상관관계

가격과 배기량

음의 상관관계

가격과 주행거리
가격과 소유자 이전 횟수

06 향후 분석 계획

엔카, KB차차차 웹크롤링을 통한 데이터 수집

앞서 보배드림 사이트를 웹크롤링할 때 가져온 차량정보들을 위 사이트에서 동일하게 수집하여,
동일한 방식으로 전처리하는 작업을 수행할 예정입니다.

기계학습 모델에 데이터 학습

수집한 중고차 데이터를 사이트별로 각각 기계학습 모델에 학습시킨 후,
각 모델의 예측 성능을 비교해볼 예정입니다.