

워싱턴 공공 자전거

수요량 예측

회귀모델 실험을 통해 모델 최적화

분석 코드는 [여기\(클릭\)](#)서 확인하실 수 있습니다.

문상혁

INDEX

01 배경과 목적

02 EDA

기술통계분석
상관관계 분석
변수별 분석

03 Feature Engineering

파생변수 생성 및 선별
이상치 처리 및 표준화

04 예측 모델 개발

다중선형회귀(정규화X, Lasso, Ridge)
다항회귀(Lasso, Ridge)
랜덤포레스트 / XGboost

04 결론



목적

- (1) 자전거 대여 패턴을 분석하여 자전거 배치 및 운영 전략 최적화
- (2) 정확한 대여 수요 예측으로 시스템 효율화 및 사용자 만족도 제고

목표

1. 머신러닝 모델 실험을 통해 최적의 수요 예측 모델을 개발한다.
2. 핵심 평가 지표인 RMSLE를 최소화한다.



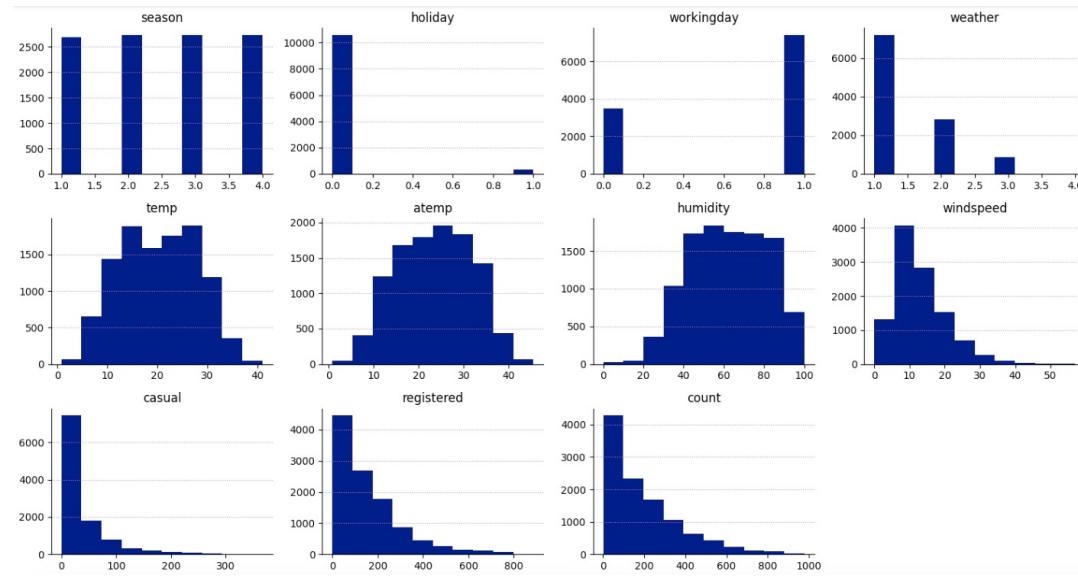
- ✓ Train 데이터 및 Test 데이터 존재, Train 데이터로 모델을 학습시키고 Test 셋의 대여량을 예측한다.

데이터 설명

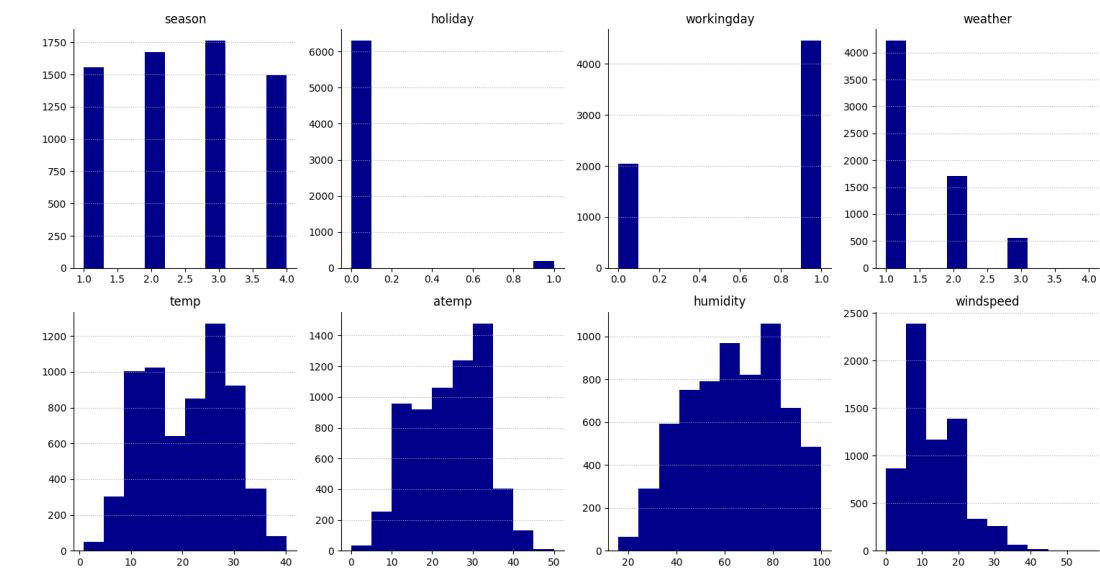
컬럼명	설명	비고
datetime	자전거 대여 기록의 날짜 및 시간	datetime
season	계절(1:봄, 2:여름, 3:가을, 4:겨울)	int
holiday	공휴일 여부 (0: 평일, 1: 공휴일)	int
workingday	근무일 여부 (0: 주말/공휴일, 1: 근무일)	int
weather	날씨 상황 (1: 맑음, 2: 구름낀/안개, 3: 약간의 비/눈, 4: 폭우/폭설)	int
temp	실측 온도 (섭씨)	float
atemp	체감 온도 (섭씨)	float
humidity	습도(%)	float
windspeed	풍속(m/s)	float
casual	등록되지 않은 사용자의 대여 수	Int / Train 셋에만 존재
registered	등록된 사용자의 대여 수	Int / Train 셋에만 존재
count	총 대여 수(종속변수)	Int / Train 셋에만 존재 / 종속변수

표 1. 데이터 컬럼별 설명

기술통계분석 – 데이터 분포 확인



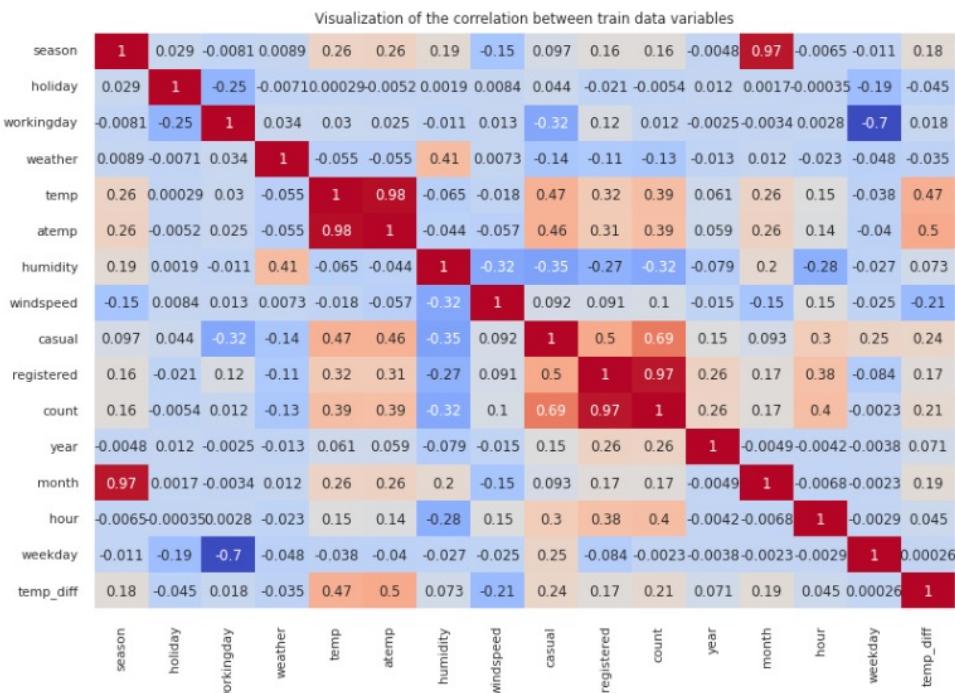
그래프 1. Train 데이터 컬럼별 분포 시각화



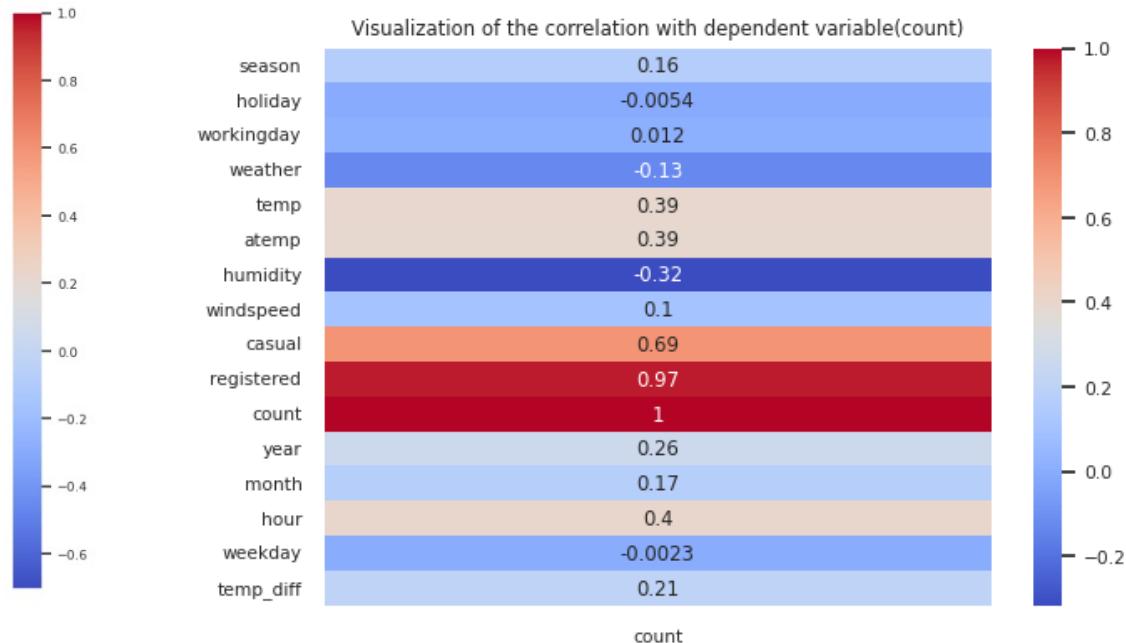
그래프 2. Test 데이터 컬럼별 분포 시각화

- Train 데이터에는 **casual**, **registered**, **count** 컬럼 존재하고, Test 데이터에는 존재하지 않으므로 모델 훈련시 **casual**, **registered**는 제외
- 범주형 변수를 제외하고 연속형 변수(**temp**, **atemp**, **humidity**, **windspeed**, **casual**, **registered**, **count**)의 분포 확인
 - ✓ **temp**, **atemp**는 중앙에 밀집된 정규분포의 형태
 - ✓ **humidity**는 오른쪽으로 약간 치우쳐진 분포
 - ✓ **windspeed**, **casual**, **registered**, **count**는 왼쪽으로 치우친 분포

상관관계 분석



그래프 3. 변수별 상관관계 시각화



그래프 4. 종속변수와의 상관관계 시각화

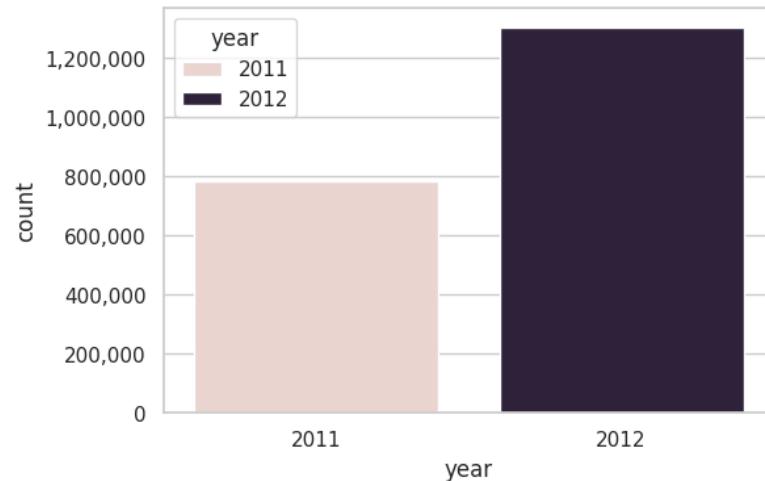


* **빨간색** 표기된 변수는 상관관계가 비교적 큰 변수

* count = casual + registered 이므로 당연히 casual, registered 두 변수와의 상관관계는 높으므로 해당 변수간에는 다중공선성이 존재할 것이다.

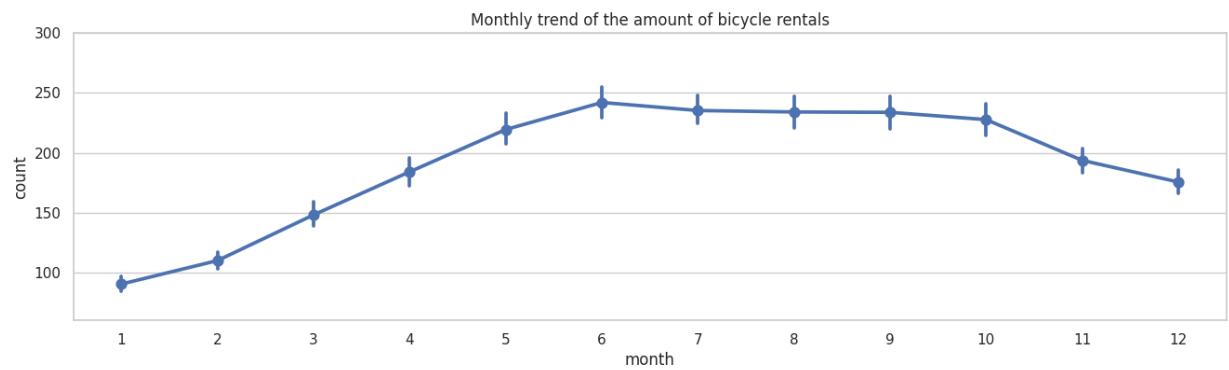
* temp, atemp는 실측온도와 체감온도로, 두 변수는 거의 동일하므로 다중공선성 존재할 것이다.

변수별 분석 – datetime 변수에서 파생하여 분석

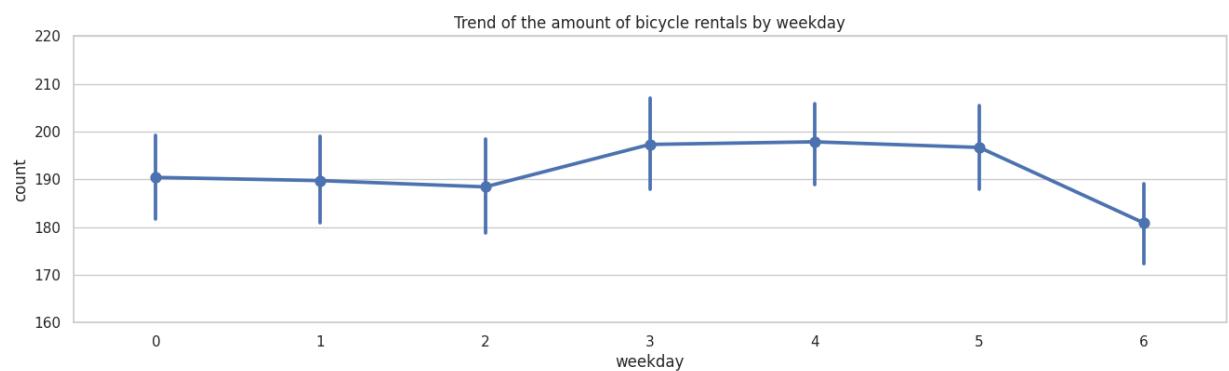


그래프 5. 연도별 자전거 대여량 차이

- ✓ 1년 사이에 평균 대여량이 증가했다.
 - 서비스 성장에 따른 자연스런 현상이라고 유추 가능
- ✓ 6월까지 대여량 상승 여름/가을인 6~10월에 대여량 최고
 - 계절별 대여량 비교하여 교차 검증해보기로 한다.
- ✓ 목, 금, 토 대여량이 다른曜일보다 높다.

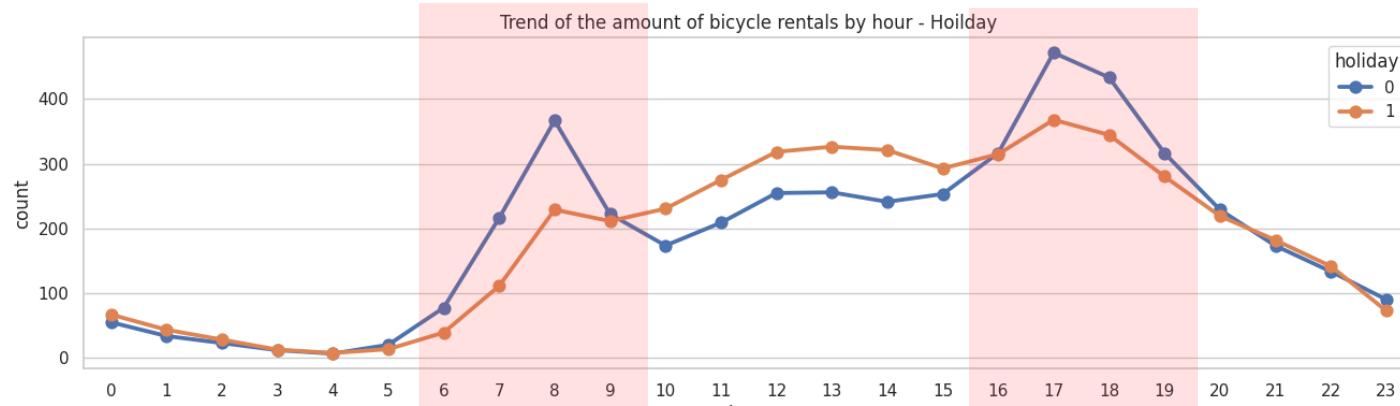


그래프 6. 월별 자전거 대여량 추이

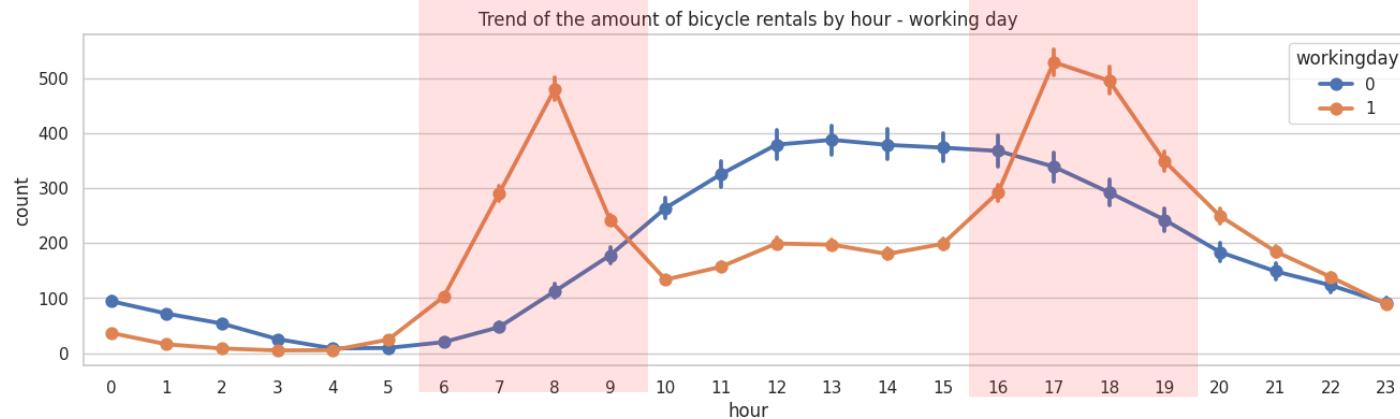


그래프 7.曜일별 자전거 대여량 추이

변수별 분석 – datetime 변수에서 시간대(hour) 파생하여 holiday, workingday 항목별 변화량 비교 분석



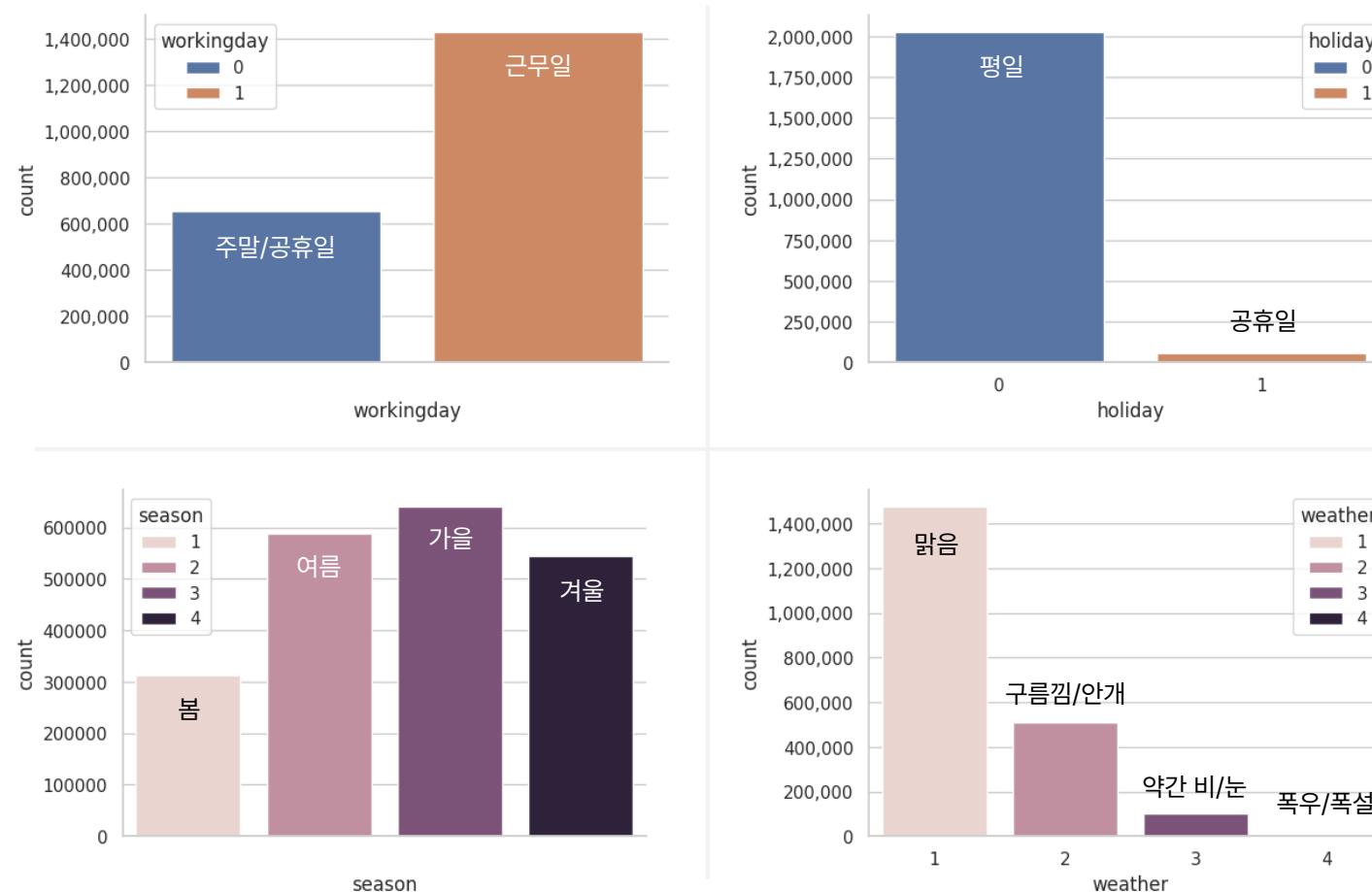
그래프 8. 시간대별 자전거 대여량 추이(공휴일 여부 기준 구분 – 0 : 평일, 1 : 공휴일)



그래프 9. 시간대별 자전거 대여량 추이(근무일 여부 기준으로 구분 – 0 : 주말/공휴일, 1 : 근무일)

- ✓ workingday, holiday 별 시간대 자전거 대여량
추이를 보니 출,퇴근 시간에 대여량이 급격히
증가하는 추세가 관찰된다.
- ✓ workingday, holiday 및 시간대 모두
대여량에 영향을 미칠 것이라는 사실 추론 가능
- ✓ workingday, holiday 별 대여량 비교하여
평일/근무일, 주말/공휴일이 대여량에 영향을
미칠 수 있다는 사실을 교차 검증해보기로 한다.

변수별 분석 – workingday, holiday, season, weather 변수의 항목별 대여량 차이 시각화



그래프 10~13. (좌측 상단부터) 근무일 여부별 / 공휴일 여부별 / 계절별 / 날씨별 자전거 대여량 차이

- ✓ 평일/근무일이 주말/공휴일보다 자전거 이용량이 많다.
→ 시계열 분석 결과 교차검증 완료
- ✓ 비나 눈이 오는 날에는 자전거 이용량이 크게 감소
- ✓ 맑은 날에 자전거를 가장 많이 대여한다.
- ✓ 봄에 가장 적게, 가을에 가장 많이 자전거를 대여
→ 시계열 분석 결과 교차검증 완료

* workingday (0: 주말/공휴일, 1: 근무일)
 * holiday (0: 평일, 1: 공휴일)
 * season (1: 봄, 2: 여름, 3: 가을, 4: 겨울)
 * weather (1: 맑음, 2: 구름낀/안개, 3: 약간 비/눈, 4: 폭우/폭설)

변수별 분석 – temp, atemp에서 temp_diff 변수 파생

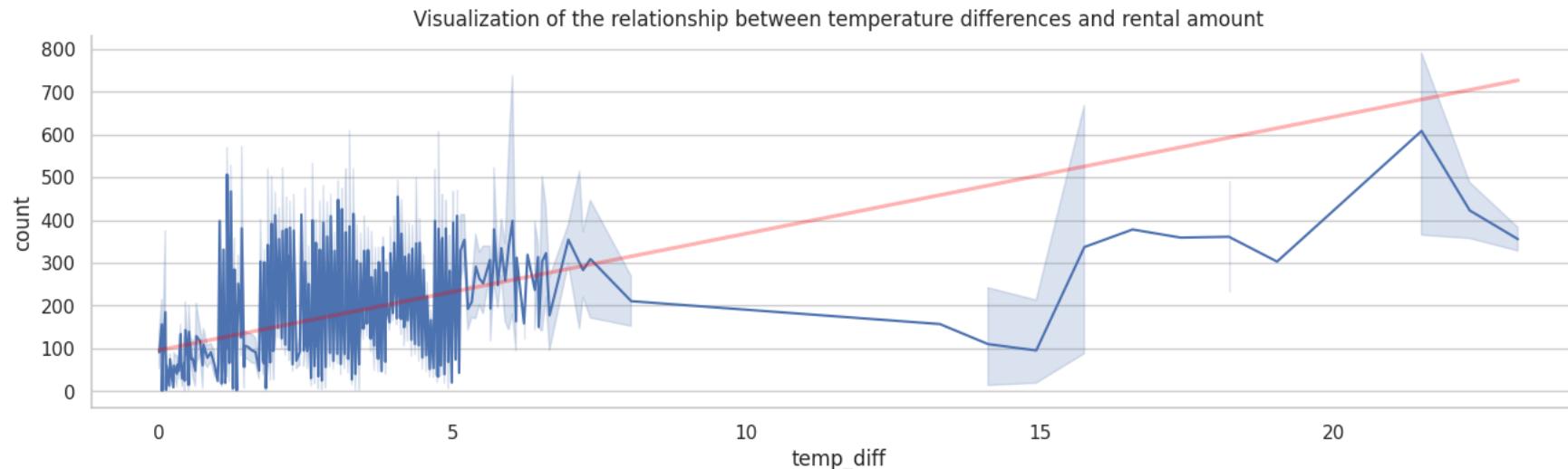


그림 14. 체감온도와 실제온도 차이와 자전거 대여량과의 관계 시각화

- 체감온도, 실제온도 간 차이에 따른 대여량 변화 확인

- 체감온도, 실제온도는 거의 동일한 추이를 가지기 때문에 두 변수를 함께 독립변수로 활용하기엔 효율이 떨어진다고 판단
- (체감온도 – 실제온도)의 절댓값으로 temp_diff라는 새로운 변수 생성

- 체감온도와 실제온도의 차이가 클수록 대여량이 증가하는 경향성 확인

- 체감온도와 실제온도 차이도 대여량에 영향을 주는 변수일 수 있다는 추론 가능

Feature Engineering 과정 요약

- ✓ 결측값과 중복값 확인 – 존재X
- ✓ datetime에서 year, month, hour, weekday 변수 파생 / temp, atemp에서 두 변수의 차이의 절댓값인 temp_diff 변수 파생
- ✓ datetime(모델 학습에 적합X), temp(atemp와 상관성 높으므로 temp_diff 변수 생성 후 삭제), casual/registered(Test 데이터에 존재하지 않으므로) 컬럼 삭제
- ✓ 이상치 처리에 관해서는 다음 장표에서 설명하기로 한다.
- ✓ 범주형 변수는 One-Hot Encoding을 통해 이진수로 변환 - season, holiday, workingday, weather, year, month, hour, weekday
- ✓ 이상치 처리 후 회귀모델 개발 시 각 모델에 맞게 변수 표준화 실시(연속형 변수만 표준화, 다향 회귀일 경우 고차항 조합에 대해서도 표준화 시행)
- ✓ 종속변수 분포의 왜도가 양수(오른쪽으로 꼬리가 길어지는 형태)이므로 로짓변환을 통해 정규성 가정을 더 잘 만족시키도록 한다.

파생변수 생성

파생변수	설명	데이터 타입
year	대여 발생 연도	int
month	대여 발생 월	int
hour	대여 발생 시간	int
weekday	대여 발생 요일(0:월, 1:화, 2:수, 3:목, 4:금, 5:토, 6:일)	int
temp_diff	체감온도와 실제온도 차이의 절댓값	float

표 2. 파생변수 변수별 설명

이상치 처리

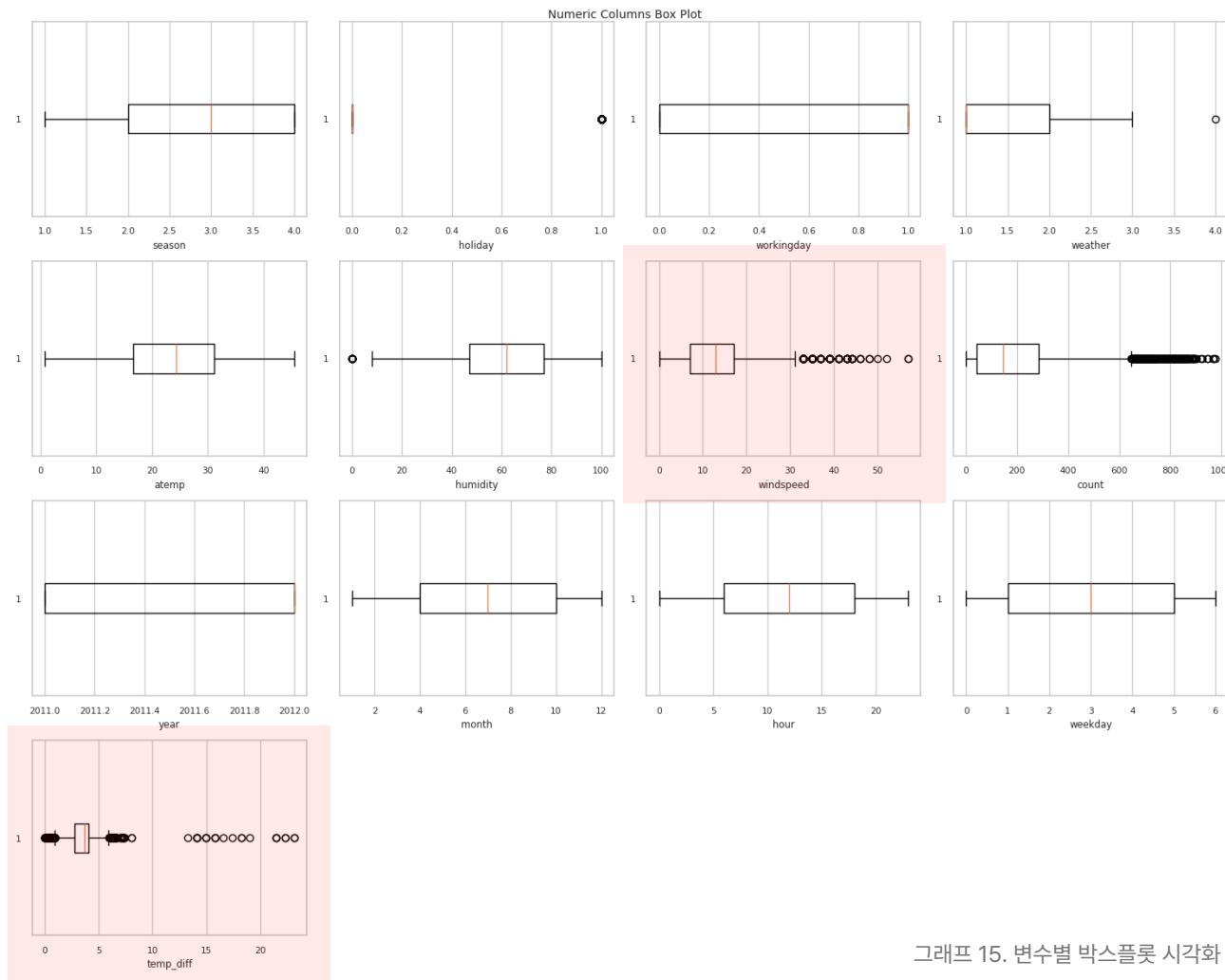


그림 15. 변수별 박스플롯 시각화

✓ temp_diff 컬럼의 이상치

데이터의 개수가 상대적으로 많고 데이터의 보존을 위해 **중앙값으로 변경하는 방식**으로 이상치를 처리한다.

✓ windspeed 컬럼의 이상치

모두 Upper 이상치이며 데이터의 보존을 위해 **최대값으로 변경하는 방식**으로 이상치를 처리한다.

✓ 범주형 변수

season, workingday, holiday, hoilday, year, month, hour는 범주형 변수이므로 이상치 처리 X

모델별 결과 요약

모델 종류	Grid-Search 여부	최적 파라미터	RMSLE	비고
다중 선형 회귀	X	-	0.596	-
다중 선형 회귀(Lasso)	O	alpha: 0.1 max_iter: 100	1.246	
다중 선형 회귀(Ridge)	O	alpha: 0.1 max_iter: 100	0.589	
다항 회귀(Lasso)	O	alpha: 0.1 max_iter: 1000 degree: 3	0.726	One-Hot Encoding 전 범주형 데이터 제외하고, 모두 표준화 적용 (고차항 포함)
다항 회귀(Ridge)	O	alpha: 1 max_iter: 1000 degree: 2	0.305	One-Hot Encoding 전 범주형 데이터 제외하고, 모두 표준화 적용 (고차항 포함)
랜덤 포레스트	O	max_depth: None min_samples_leaf: 1 min_samples_split: 2 n_estimators: 200	0.434	-
XGboost	O	learning_rate: 0.2 max_depth: 7 n_estimators: 1000 reg_alpha: 1 reg_lambda: 100	0.339	Grid Search로 찾은 최적 파라미터에서 과적합 문제 발생하여 파라미터 수정

표 3. 모델별 결과 요약

1) 제언 및 시사점

- **최적 모델 선정:** 여러 머신러닝 모델을 실험한 결과, **다항회귀(Ridge)** 모델이 가장 낮은 **RMSLE(0.30)**를 기록하여 가장 효과적인 모델로 평가됨
- **선형 모델의 활용 가능성:** 일반적인 선형회귀 모델보다 Ridge 및 다항회귀 모델이 더 나은 성능을 보였으며, 적절한 정규화와 변환을 적용하면 선형 모델도 강력한 예측 도구가 될 수 있음
- **데이터 기반 정책 수립 가능성:** 자전거 대여량을 예측하는 모델을 활용하여 수요가 많은 시간대 및 계절별 자전거 배치 최적화, 대여소 증설시 적정 배치량 산출하여 계획을 수립할 수 있음

2) 분석의 한계점

- **데이터의 한정성:** 특정 도시(워싱턴)의 공공 자전거 데이터를 기반으로 했기 때문에 다른 지역에서는 동일한 모델 성능이 보장되지 않음
- **이상치 처리의 영향:** windspeed와 temp_diff 변수의 이상치를 대체했지만, 보다 정교한 이상치 탐지 기법(ex. Isolation Forest, DBSCAN) 적용이 필요할 수 있음
- **추가적인 변수 고려 필요:** 현재 분석에서는 기온, 습도, 바람 등의 기상 요소와 날짜·시간 정보를 활용했지만, 실제 교통량, 대중교통 이용 데이터, 이벤트 일정 등의 외부 데이터를 추가하면 예측 정확도를 높일 수 있음

3) 회고 및 향후 발전 방향

- 다양한 머신러닝 모델을 실험하며 선형 모델이 특정 조건에서 강력한 성능을 발휘할 수 있음을 확인했음
- 변수 선택과 데이터 전처리 과정이 모델 성능에 미치는 영향을 경험하며, 단순한 데이터 입력이 아닌 특성 엔지니어링이 중요하다는 점을 배움
- 추가적인 변수 검토가 필요하며, 특히 외부 데이터(대중교통 이용량, 이벤트, 교통체증 등)와의 결합을 통해 모델 성능을 더욱 향상할 가능성이 있음

END

감사합니다.

문상혁