**2020. 12. 13**

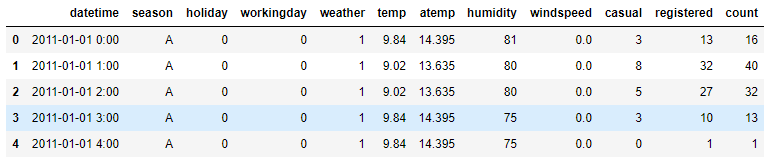
**1. 기계학습 1.2 회귀**

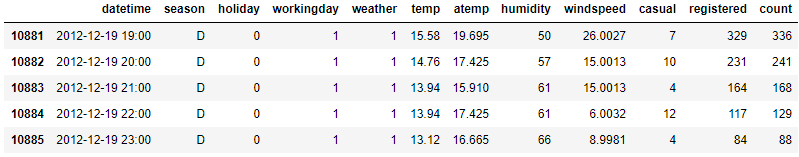
**문제 1-1**

시작 : 18:30

**1. 데이터 구조 확인**

□ head, tail을 통해 데이터의 처음과 마지막 5행을 확인

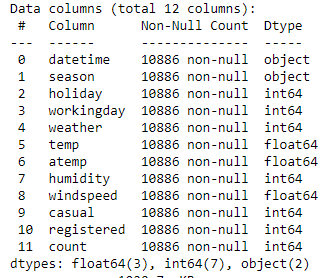




□ shape로 데이터 구조 확인 : 10886행 x 12열

□ info로 전체 컬럼명과 데이터 타입 확인

○ 12개 컬럼 중 float64 타입 3개, int64 타입 7개, object 타입 2개로 구성



□ 데이터 타입 보정

○ 컬럼 특성에 맞게 데이터 타입을 보정해 주었음

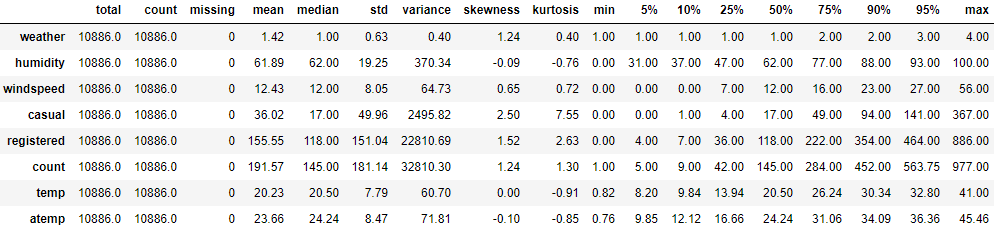
**2. DQ Check(빈도분석, 분포 분석)**

□ 연속형 변수

○ 연속형 변수들의 기초통계량과 분포분석을 진행함

- 분포가 90% 이상이 동일한 변수는 모델링 시 설명력이 떨어진다고 판단하여 제외하기로 함

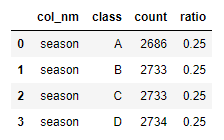
- 이 데이터에서는 모든 변수들이 고른 분포의 데이터를 가지고 있음



□ 범주형 변수

○ 범주형 변수들의 빈도분석과 분포분석을 진행함

- 이 데이터에서는 모든 변수들이 고른 분포의 데이터를 가지고 있음

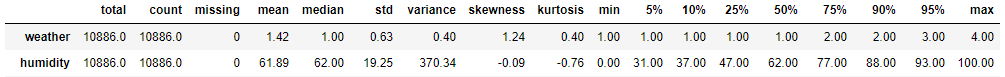


**3. 전처리(중복값, 결측값, 이상치 처리)**

**1) 중복값**

□ 중복값 확인(duplicated)

○ ~~ 사유로 중복이 발생한 것으로 생각됨



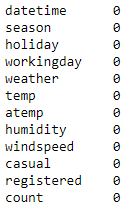
□ 중복값 제거(drop\_duplicates)

○ ~~ 사유로 중복값은 마지막 데이터만 남기기로 함

**2) 결측값**

□ 결측값 확인

○ ~~와 @@에 결측값이 n개 씩 있음을 확인함



□ 결측값 처리

○ ~~사유로 결측값은 0으로 대체함 / 결측값이 있는 데이터는 제거함

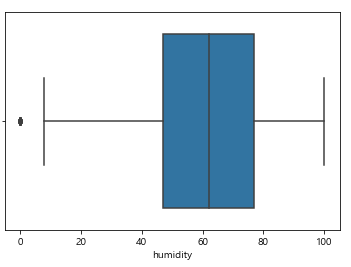
**3) 이상치**

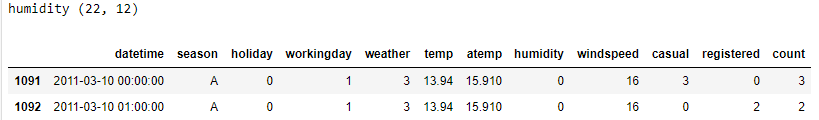
□ 이상치 확인

○ 각 컬럼별로 boxplot을 그려 IQR 방식으로 이상치를 확인함

○ 본 보고서에는 일부 변수에 대해서만 그래프를 첨부함

- humidity 변수에서는 총 22개의 이상치를 확인함





□ 이상치 처리

○ 통계적으로 이상치이기는 하나 현실적으로 발생할 수 있는 경우이기 때문에 min/max값 경계 밖의 데이터를 min/max값으로 보정함

또는 ○ min/max값 경계 밖의 데이터는 제거함 : 제거 후 데이터는 10,864건

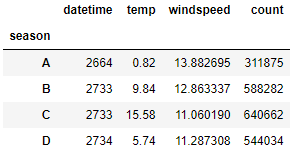
또는 ○ 통계적으로 이상치이기는 하나 현실적으로 발생할 수 있는 경우이기 때문에 유지함

**4. 요약데이터로 변환(groupby)**

□ season 변수를 기준으로 요약데이터를 생성함

○ datetime의 count

○ temp의 최소값



**5. 파생변수 생성**

□ RFM의 요인을 기준으로 파생변수를 생성함 / 아니면 datetime에서 연월일, 시간대 등

○ Recency(최근성) : datetime이 3000일 이상이면 1, 2000일 이상이면 2, 그 외는 3

○ Frequency(빈도) : ID별 거래 횟수(ID의 count가 10 이내면 1, 20 이내면 2, 그 외는 3

○ Monetary(거래규모) : count와 temp의 곱(교호작용의 효과를 확인하고자 한다)

O 회당 값 : ~~을 파악하기 위해 temp를 count로 나눔



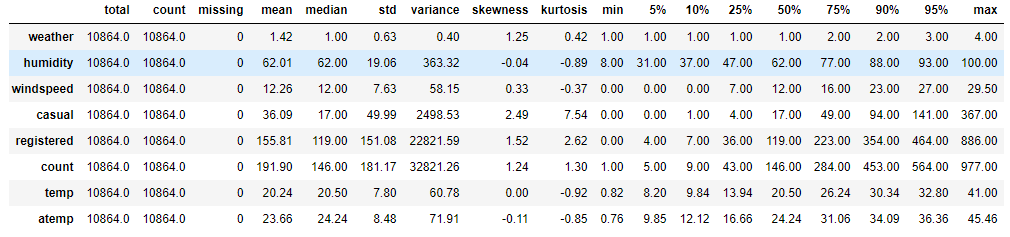
**6. 데이터 마트 DQ Check, 변수선택 및 EDA**

□ 데이터 전처리 후 마트 DQ Check 재수행함

○ 연속형 변수 분포분석

- 분포가 90%이상이 동일한 변수는 모델링 시 설명력이 떨어진다고 판단하여 제외하기로 함

- 이 데이터에서는 모든 변수들이 고른 분포의 데이터를 가지고 있음



□ 데이터 전처리 후 마트 DQ Check 재수행함

○ 범주형 변수 분포분석

- 한 구간의 분포가 90% 이상인 변수는 모델링 시 설명력이 떨어진다고 판단하여 제외 필요

- Frequency가 class 3에 91%가 집중되어 있어 동 변수를 제외하기로 함

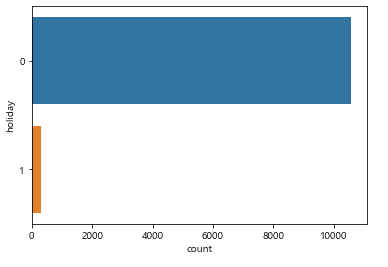
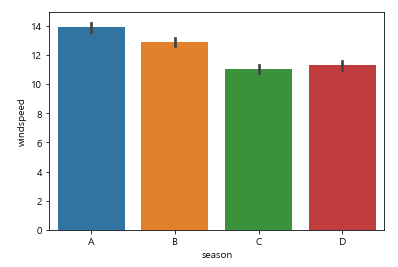


□ EDA(시각화)

○ 변수별 시각화를 통해 각 변수의 특성을 분석하였음

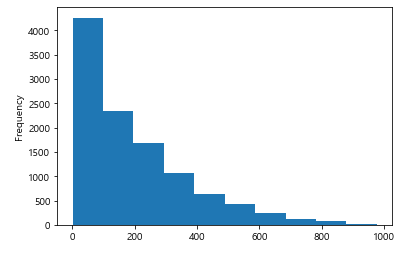
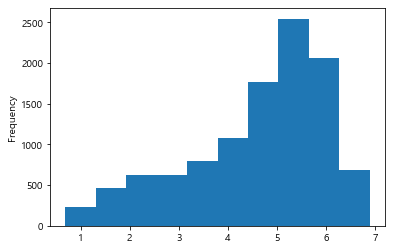
- season class에 따라 windspeed의 평균에 차이가 있었음을 확인할 수 있었음

- holiday의 빈도는 0이 압도적으로 많았음



**7. 종속변수 분포 확인 및 전처리**

□ 종속변수의 분포가 한 쪽으로 치우쳐져 있어 log값을 취함

🡪 

**8. 변수 더미화**

□ 특정 알고리즘을 활용한 모델링 과정에서의 에러 발생을 방지하기 위해 범주형 변수에 대해 더미화를 진행하였음

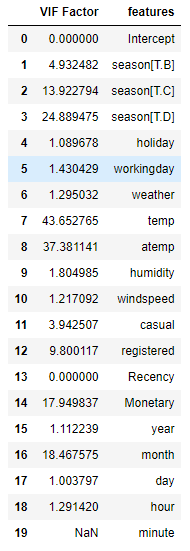
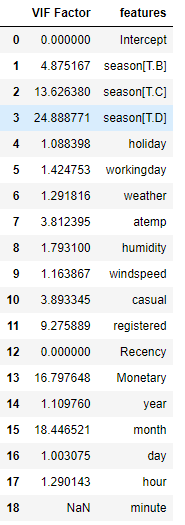
**9. 다중공선성 제거**

□ 다중공선성은 예측변수 사이에 높은 상관관계가 있을 때 발생하며, 회귀계수 추정치의 신뢰성과 안정성에 문제를 발생시키므로 VIF(분산팽창지수)를 활용하여 사전 제거함

○ temp와 atemp의 VIF Factor 값이 각각 43.7과 37.4로 다중공선성 발생이 관찰되어 둘 중 temp 변수를 제거함

○ temp 제거후 atemp의 VIF Factor 값이 3.8로 안정화 된 것을 확인함

○ Monetary 변수의 값이 16.8로 10을 초과하나 데이터 분석상 필요할 것으 판단되어 유지하여 모델에 사용하기로 함

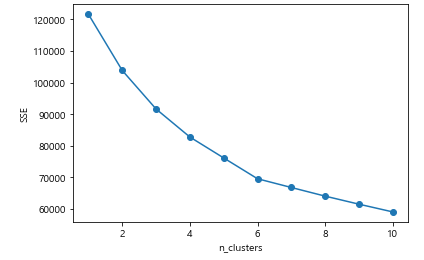
 

**10. Train, Test set 분리**

□ 데이터 마트를 Train, Test set 7:3 비율로 분리함

**11. 데이터 정규화**

□ StandardScaler로 X\_train과 X\_test를 정규화함

**13. 군집화**

□ K-means clustering을 통해 군집 수 2~10개로 변화시켜가며 오차제곱합(SSE)의 변화와 실루엣계수를 통해 최적군집수를 찾아보았음

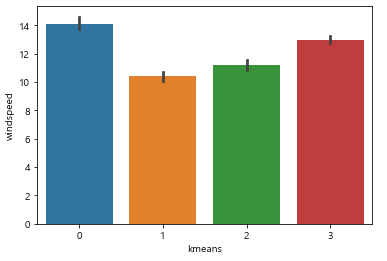
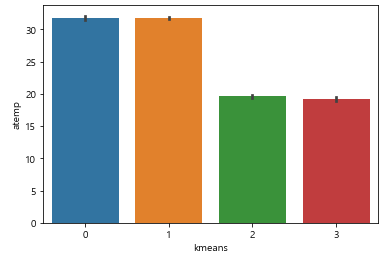
○ 군집 개수가 4개일 때 최적인 것으로 판단됨

**14. 군집화 결과 프로파일링**

□ 군집별 통계적 특성을 확인해 보았음(본 보고서에는 일부 그래프만 첨부)

○ 군집 0과 1의 atemp의 평균은 30을 상회하나 군집 2와 3은 20을 밑돌았음

○ 군집 0의 windspeed의 평균은 ~~



**15. 데이터마트에 군집화 결과 활용**

□ X\_train과 X\_test에 대한 군집화를 수행한 후 군집화 결과를 데이터 마트에 더하여 모델 구축에 활용하기로 함

**16. 분류 모델링(기본 모델)**

□ 다양한 알고리즘을 활용하여 기본 모델을 구축함

○ Ridge

○ Lasso

○ ElasticNet

○ Huber Regression

○ Decision Tree Regression

○ RandomForest Regression

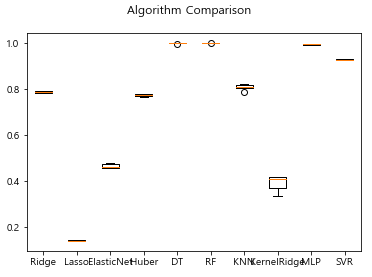
○ K Neighbors Regression

○ KernelRidge

○ MLPRegressor

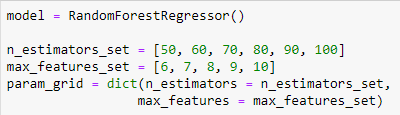
○ SVR

□ Kfold=5 Cross validation을 수행한 결과 RandomForest Regresstion의 score가 1.00으로 가장 높았음



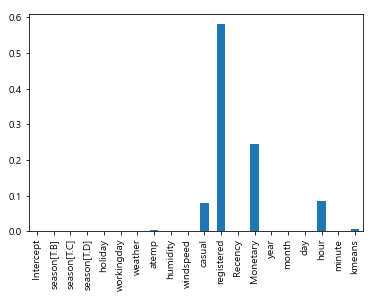
**17. 파라미터 튜닝 및 최종 모델 선정**

□ 앞서 선택한 RandomForest Regression의 파라미터를 Gridsearch CV로 튜닝하여 성능을 개선함



□ best parameter로 max\_feature : 10 , n\_estimators : 100이 선정됨

□ 본 모델의 변수중요도는 xx, xx 순으로 높았음



**18. Test set 활용하여 예측 수행 및 성능평가**

□ R square는 0.99로 모델의 설명력이 매우 우수하다고 볼 수 있음

- 결정계수 라고 하며, R2 으로도 표현된다. 용처는 회귀모형의 설명력을 표현하는 것. P-value와 같이 0과 1 사이의 값으로 나타나는데, 0에 가까울수록 설명력이 낮고, 1에 가까울수록 높다고 해석할 수 있다.

□ MSE

- MSE가 작은 모형일수록 회귀식의 오차가 줄기때문에 그만큼 현상을 잘 설명한다고 할 수 있습니다.

□ RMSE

- 평균 제곱근 오차(Root Mean Square Error; RMSE)는 추정 값 또는 모델이 예측한 값과 실제 환경에서 관찰되는 값의 차이를 다룰 때 흔히 사용하는 측도

□ MAE

- 절대값(에러 - 실제값) 의 평균인 에러 지표. 에러의 절대 값 그 자체를 나타내기 때문에, 값은 낮을수록 좋다.

□ MAPE

- 평균 절대 백분율 오차(MAPE)는 정확도를 오차의 백분율로 표시합니다. MAPE는 백분율이기 때문에 다른 정확도 측도 통계량보다 더 쉽게 이해할 수 있습니다. 예를 들어 MAPE가 5이면 예측 값은 평균 5% 벗어납니다.