O 사전 세팅

1. [워드] 단락 뒤 공백 제거
2. 캡처도구 켜두기
3. C:\windows\fonts에서 맑은 고딕 파일(malgun.ttf) Jupyter Notebook으로 옮겨두기  
   (fonts 폴더에서 ‘맑은’ 검색하면 됨)
4. 미리 코드 쳐두기(라이브러리 등)
5. 문제 파일 확인

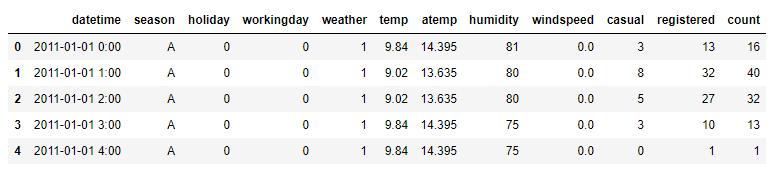
O 서식

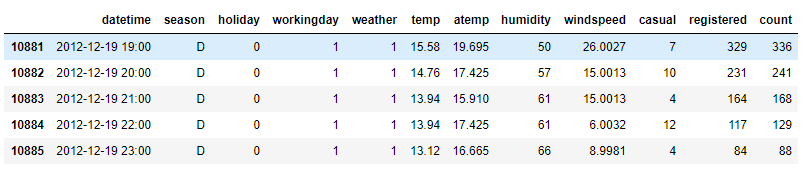
1. **대분류(Bold, 14pt)**
   1. 중분류
      1. 소분류

**2021. 3. 22**

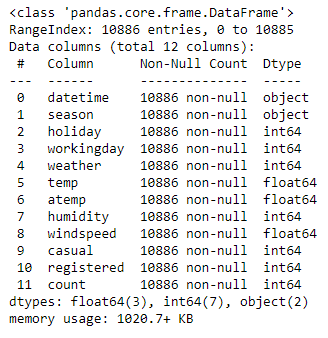
**1. 기계학습 1.2 회귀**

1. **데이터 구조 확인 및 EDA**
   1. head, tail을 통해 데이터 처음과 마지막 5행 확인

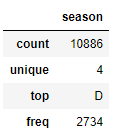
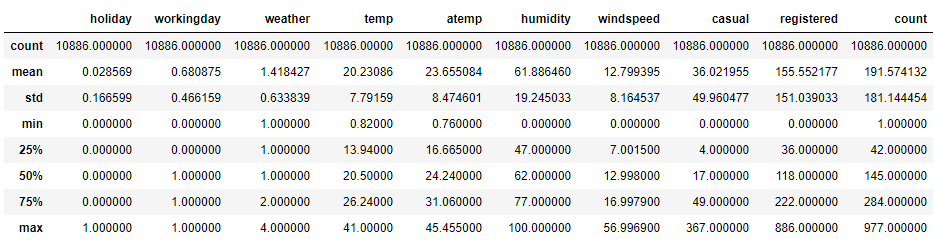




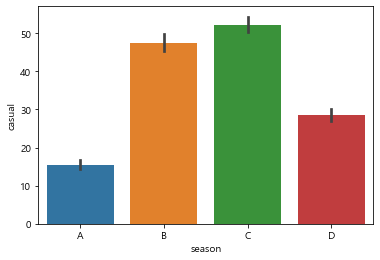
* 1. shape로 데이터 구조 확인 : 10,886행 x 12열
  2. info로 컬럼별 데이터 타입 확인



* 1. describe로 빈도와 분포 확인



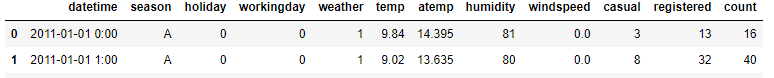
* 1. 시각화를 통한 데이터 구조 확인  
     (전체 컬럼 중 몇 가지 컬럼만 선별하여 첨부함)
     1. Season class에 따라 casual 평균 값에 차이가 있음을 확인했으며, season이 C일 경우 casual의 평균은 50이 넘어 전체 season 중 가장 높았으며, 반대로 A의 경우 약 15로 가장 낮았음



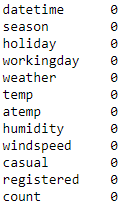
* + 1. Season별 빈도는 모두 2500 정도로 고르게 분포하고 있음



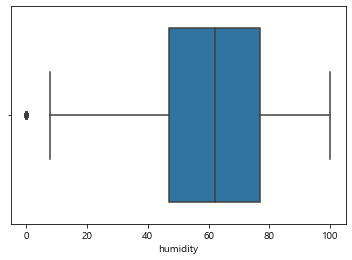
1. **데이터 마트 DQ check 및 변수 선택 @@전 버전 ‘분류’ 참고**
2. **데이터 전처리**
   1. 중복값 처리
      1. ~~ 사유로 중복이 발생한 것으로 생각됨

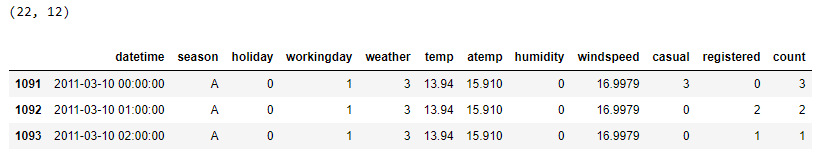


* + 1. ~~ 하여 중복값은 마지막 데이터만 남기기로 함
  1. 결측값 처리
     1. ~~와 @@에 결측값이 n개 있음을 확인함



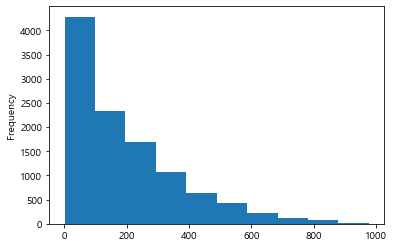
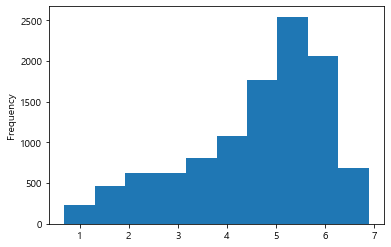
* + 1. ~~ 사유로 결측값은 0으로 대체함 / 결측값이 있는 데이터는 제거함 / 중앙값
  1. 이상치 처리
     1. 각 컬럼별로 boxplot을 그려 IQR 방식으로 이상치를 확인함
     2. 본 보고서에는 일부 변수에 대해서만 plotting함
        1. humidity 변수에서는 총 22개의 이상치를 확인함



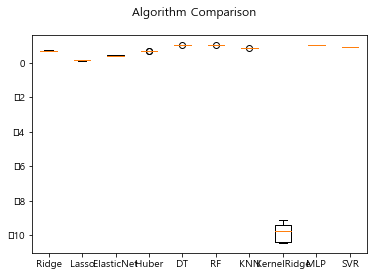
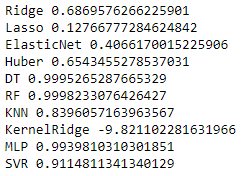


* + 1. 데이터를 최대한 활용하기 위해 min/max값 경계 밖의 데이터를 min/max로 보정함 / 제거함

1. **파생변수 생성 @@전 버전 ‘분류’ 참고**
2. **종속변수 분포 확인**
   1. 종속변수의 분포가 한 쪽으로 치우쳐져 있어 log를 취함

** 🡪 **

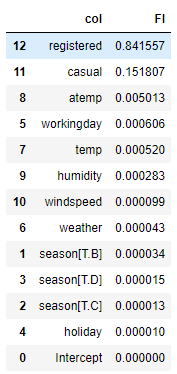
1. **변수 더미화**
   1. 특정 알고리즘을 활용한 모델링 과정에서의 에러 발생을 방지하기 위해 범주형 변수에 대해 더미화를 진행하였음
2. **다중공선성 제거 @@전 버전 ‘분류’ 참고**
3. **Train, Test set 분리**
   1. 데이터 마트를 Train, Test set 7:3 비율로 분리함
4. **데이터 정규화**
   1. Standard Scaler로 데이터를 정규화함
5. **군집화를 통한 파생변수 생성 @@전 버전 ‘분류’ 참고**
6. **모델링 및 기초 모델 선정 @@XGB는 필요시 전 버전 참고**
   1. 다양한 알고리즘을 활용하여 기본 모델을 구축함
      1. Ridge
      2. Lasso
      3. ElasticNet
      4. Huber Regression
      5. Decision Tree Regression
      6. RandomForest Regression
      7. K Neighbors Regression
      8. KernelRidge
      9. MLPRegressor
      10. SVR
   2. Kfold=5 Cross validation을 수행한 결과 RandomForest Regresstion의 score가 1.00으로 가장 높았음



1. **파라미터 튜닝 @@ RF 외 GridsearchCV는 전 버전 참고**
   1. 앞서 선택한 RandomForest Clasifier의 파라미터를 Gridsearch CV로 튜닝하여 성능을 개선함



* 1. best parameter로 max\_feature : 10, n\_estimators : 90이 선정됨
  2. 본 모델의 변수중요도는 registered, casual 순으로 높았음

1. **Test set 활용하여 예측 수행 및 성능평가**
   1. R square는 0.99로 모델의 설명력이 매우 우수하다고 볼 수 있음
      1. 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.
   2. MSE
      1. MSE가 작은 모형일수록 회귀식의 오차가 줄기때문에 그만큼 현상을 잘 설명한다고 할 수 있습니다.
   3. RMSE
      1. 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
   4. MAE
      1. 절대값(에러 - 실제값) 의 평균인 에러 지표. 에러의 절대 값 그 자체를 나타내기 때문에, 값은 낮을수록 좋다.
   5. MAPE
      1. 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.