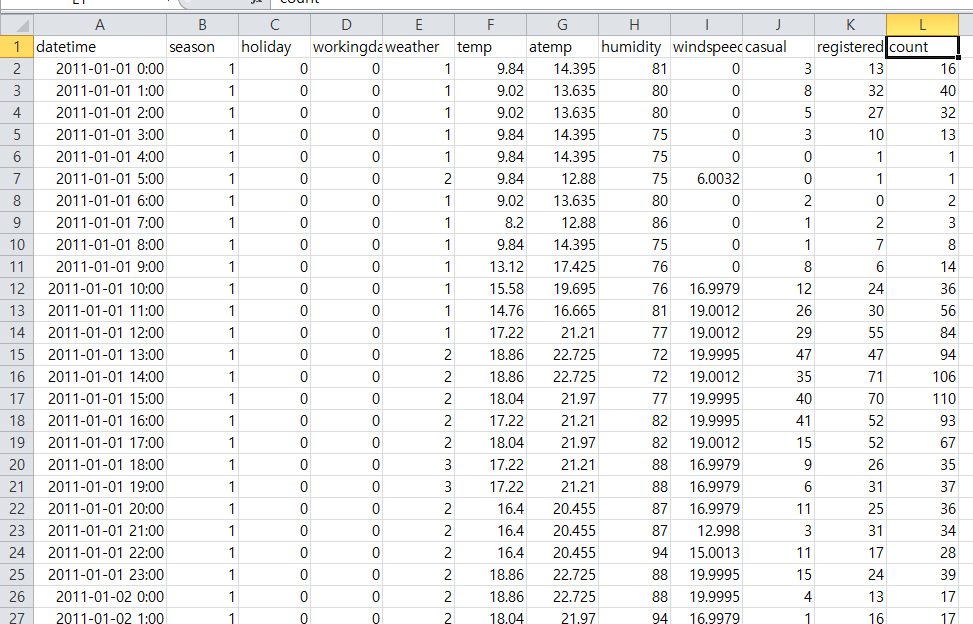
<Quiz 13>

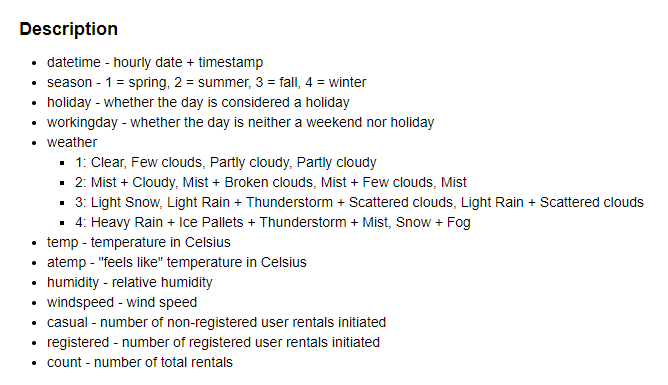
이번 시간에는 Bike sharing demand 데이터(이하 bike 데이터)를 이용하고자 한다. 이번 시간에는 다소 low level의 feature들을 다룬다. 이를 위해 간단한 feature engineering를 수행하고자 한다. 또한, 기초적인 EDA(탐색적 데이터 분석)을 수행해볼 것이다. 마지막으로는 자전거 공유 시스템의 수요를 예측할 수 있는 회귀모델을 구축하고자 한다.

Bike Sharing Demand 데이터셋은 워싱턴 DC의 Capital Bikeshare 프로그램에서 자전거 대여 수요를 예측하는 Kaggle Competition에서 이용되었던 데이터셋이다. (https://www.kaggle.com/c/bike-sharing-demand)

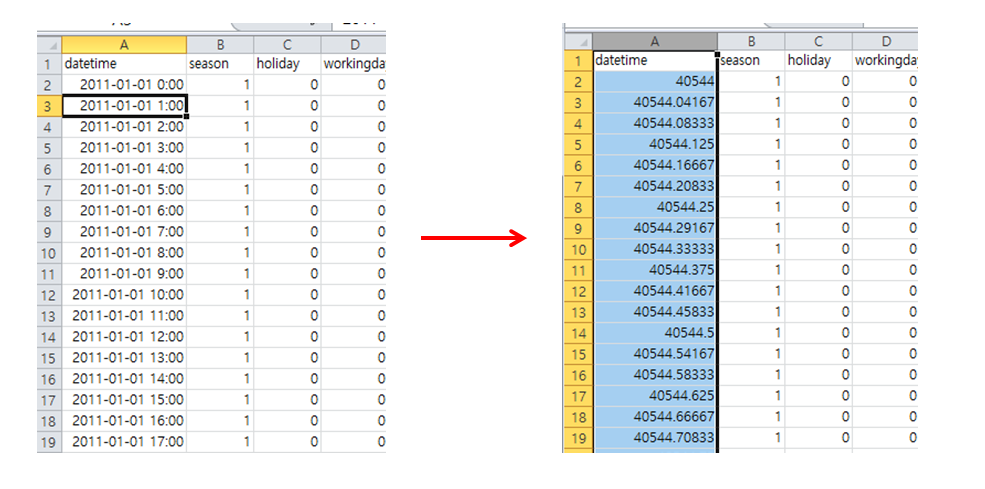
우선 train.csv 파일을 열어보면 다음과 같은 형태로 데이터셋이 주어진 것을 알 수 있다.



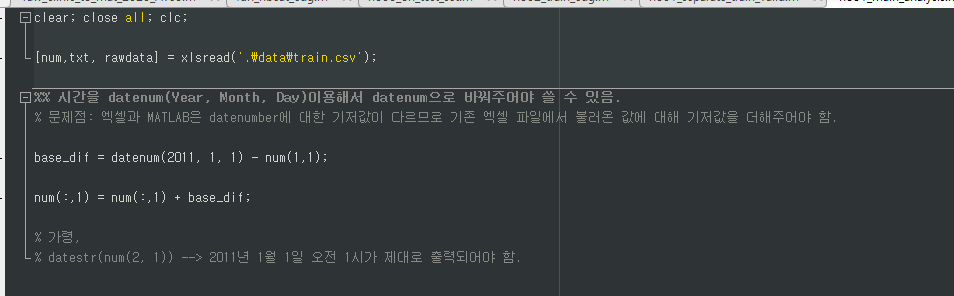
우리가 feature로 사용할 수 있는 데이터는 Datetime(날짜), season(계절. 1: 봄, 2: 여름, 3: 가을, 4: 겨울), holiday(국경일 여부), workingday(주중 여부), weather(날씨), temp(온도), atemp(체감온도), humidity(습도), windspeed(풍속)까지 이다. 뒤의 casual, registered, count는 해당 날짜에 자전거를 빌려간 횟수에 해당된다. casual은 비회원 대여 횟수, registered는 회원 대여 횟수, count는 비회원, 회원 대여 횟수를 합친 값을 의미한다.



1. 엑셀을 이용한 feature engineering. MATLAB에서 시간 데이터를 쉽게 불러올 수 있게 하기 위해 셀 서식 조정을 통해 아래와 같이 datetime을 숫자형식으로 수정해주도록 하자.

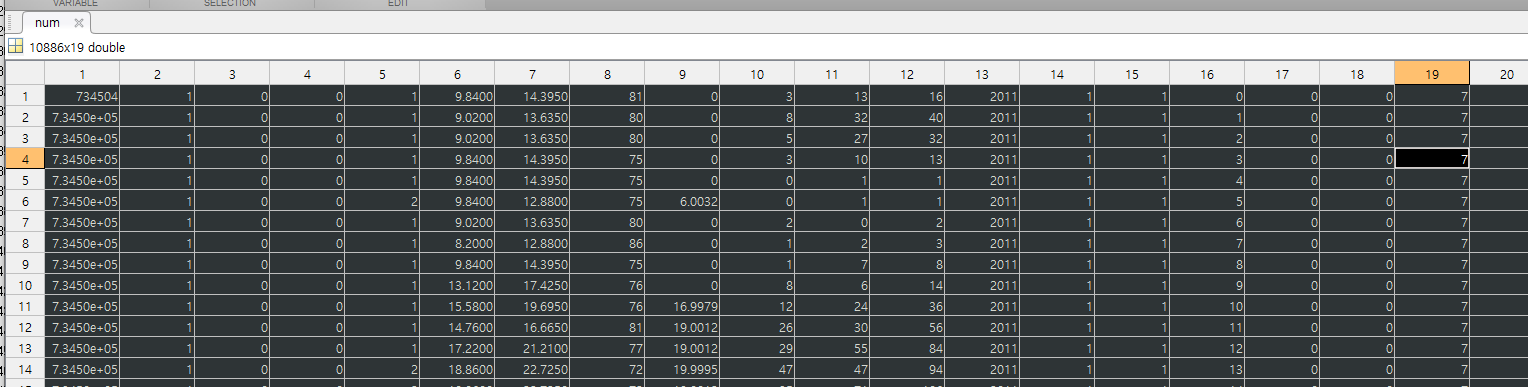


2. 그런 다음 MATLAB에서 xlsread() 함수를 이용해 데이터셋을 불러온 뒤, 아래의 코드를 이용해 MATLAB에서 시간 데이터를 이용할 수 있도록 기저 시간을 수정해주도록 하자.



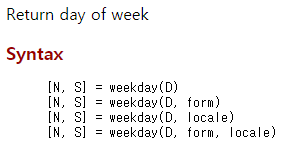
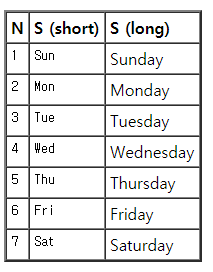
3. 연도, 월, 일, 시간, 분, 초를 나눠서 새로운 feature로 만들어주도록 하자.  
(힌트: 사용 가능한 함수는 datestr()와 datevec()이다.)

아래의 그림에서 13~18번 열은 각각 연, 월, 일, 시, 분, 초를 의미한다.



또, weekday() 함수를 이용해 요일을 추출하도록 하여 19번 열에 넣도록 하자.

Weekday() 함수의 syntax 및 출력 결과는 다음과 같다.

, 

4. 여러 가지 시간 단위에 따른 대여량의 차이를 확인해보도록 하자.



이 결과를 보면 2011년에 비해 2012년에 대여량이 더 많았음을 알 수 있다. 겨울에 대여량이 감소하는 것 또한 알 수 있다. 일별 대여량은 크게 다르지 않는데, 20일 이후 데이터들은 삭제되어 있음을 알 수 있다. 그 이유는 train data에는 1~20일까지의 데이터가 들어있는 반면, test data에는 21~31일 까지의 데이터가 들어있기 때문이다. 또, 시간별 대여량을 보면 늦은 밤과 새벽 시간에는 대여량이 적은 것을 확인할 수도 있다.

5. 여러 조건에 따른 시간 별 count를 비교해보도록 하자.

5-1. 시간 별 count를 모든 조건에 대해 그려보자. Error bar는 1.96\*SEM으로 표기하여 95% 신뢰구간을 표현해줄 것 (이하 그래프에 있는 모든 error bar는 동일한 방법으로 처리할 것).



5-2. 시간 별 count를 working day 여부에 따라 다른 그래프로 그려보자.



이 결과를 보면 working day에는 출퇴근 시간에 대여량이 많은 반면, non-working day에는 점심 시간대에 대여량이 많은 것을 알 수 있다.

5-3. 요일수(일요일-토요일)에 따른 대여량을 확인해보자.



이 결과를 보면 토요일, 일요일에는 점심시간 주변에서 대여량이 많은 반면, 주중에는 출퇴근 시간에 대여량이 많은 것을 알 수 있다.

5-4. 날씨에 따른 대여량을 확인해보자.



이 결과를 보면 맑은 날에 가장 대여량이 많았으며, 그 다음으로 안개 낀 날, 눈 온 날, 폭우 쏟아진 날 순서대로 대여량이 줄어들었음을 알 수 있다.

5-5. 계절 별로 대여량을 확인해보자.



이 결과를 보며 겨울에 대여량이 가장 적은 것을 알 수 있다.

6. 이제는 Regression 모델들을 이용해 직접 학습과 테스트를 수행해보고자 한다.

\* 학습 세트와 훈련 세트에 대해 설명할 것.

\* Regression Learner 이용해 학습하는 과정에 대해 설명할 것.

- 이용하면 안되는 feature 골라내기

- PCA에 대해 전반적인 설명 할 것

\* k-fold CV에 대한 설명

