**Dataset**[**¶**](#gjdgxs)

Dataset is about bike rental service. It contains bike sharing counts aggregated on hourly basis. Total number of records: 17379 hours (2 years of rental services). There are following features:

* instant: record index
* dteday : date of record
* hr : hour (0 to 23) of record
* season : season (1:springer, 2:summer, 3:fall, 4:winter)
* yr : year (0: 2011, 1:2012)
* mnth : month ( 1 to 12)
* holiday : whether day is holiday or not (extracted from <http://dchr.dc.gov/page/holiday-schedule>)
* weekday : day of the week
* workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
* weathersit : weather situation in the current timeframe
  + 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  + 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  + 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  + 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
* temp : Normalized temperature in Celsius. The values are divided to 41 (max)
* atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
* hum: Normalized humidity. The values are divided to 100 (max)
* windspeed: Normalized wind speed. The values are divided to 67 (max)
* casual: count of casual users (target 1)
* registered: count of registered users (target 2)
* cnt: count of total rental bikes including both casual and registered (target total)

In [1]:

**import** **pandas** **as** **pd**

In [2]:

pd.read\_csv('data/hour.csv')

Out[2]:

|  | **instant** | **dteday** | **season** | **yr** | **mnth** | **hr** | **holiday** | **weekday** | **workingday** | **weathersit** | **temp** | **atemp** | **hum** | **windspeed** | **casual** | **registered** | **cnt** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1 | 2011-01-01 | 1 | 0 | 1 | 0 | 0 | 6 | 0 | 1 | 0.24 | 0.2879 | 0.81 | 0.0000 | 3 | 13 | 16 |
| **1** | 2 | 2011-01-01 | 1 | 0 | 1 | 1 | 0 | 6 | 0 | 1 | 0.22 | 0.2727 | 0.80 | 0.0000 | 8 | 32 | 40 |
| **2** | 3 | 2011-01-01 | 1 | 0 | 1 | 2 | 0 | 6 | 0 | 1 | 0.22 | 0.2727 | 0.80 | 0.0000 | 5 | 27 | 32 |
| **3** | 4 | 2011-01-01 | 1 | 0 | 1 | 3 | 0 | 6 | 0 | 1 | 0.24 | 0.2879 | 0.75 | 0.0000 | 3 | 10 | 13 |
| **4** | 5 | 2011-01-01 | 1 | 0 | 1 | 4 | 0 | 6 | 0 | 1 | 0.24 | 0.2879 | 0.75 | 0.0000 | 0 | 1 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **17374** | 17375 | 2012-12-31 | 1 | 1 | 12 | 19 | 0 | 1 | 1 | 2 | 0.26 | 0.2576 | 0.60 | 0.1642 | 11 | 108 | 119 |
| **17375** | 17376 | 2012-12-31 | 1 | 1 | 12 | 20 | 0 | 1 | 1 | 2 | 0.26 | 0.2576 | 0.60 | 0.1642 | 8 | 81 | 89 |
| **17376** | 17377 | 2012-12-31 | 1 | 1 | 12 | 21 | 0 | 1 | 1 | 1 | 0.26 | 0.2576 | 0.60 | 0.1642 | 7 | 83 | 90 |
| **17377** | 17378 | 2012-12-31 | 1 | 1 | 12 | 22 | 0 | 1 | 1 | 1 | 0.26 | 0.2727 | 0.56 | 0.1343 | 13 | 48 | 61 |
| **17378** | 17379 | 2012-12-31 | 1 | 1 | 12 | 23 | 0 | 1 | 1 | 1 | 0.26 | 0.2727 | 0.65 | 0.1343 | 12 | 37 | 49 |

17379 rows × 17 columns

**Task**[**¶**](#30j0zll)

* Preprocessing (1 point)
* Feature engineering (2 points)
* Use different advanced regression algorithms:
  + KNN Regressor (1 point)
  + Decision Tree Regressor (1 point)
  + Random Forest / Extra Trees Regressor (1 point)
  + SVM Regressor (1 point)
  + Boosting Regressor - any Boosting algorithm with Regression (1 point) NB! The task is considered done for each model if there is hyperparameters tuning and validation is used.
* Build regression model for each year in question (2 points):
  + Target is cnt, exclude casual and registered from the model
  + Validate either on last 7 days of each months or on random subsample
  + Check the quality of prediction
  + Build regression model for the first year and make predictions about the second. The results are better or worse? Why? (1 points)
* Build regression model for each year in question (2 points):
  + Multitarget: casual and registered, exclude cnt from the model
  + Validate either on last 7 days of each months or on random subsample
  + Check the quality of prediction
  + Build regression model for the first year and make predictions about the second. The results are better or worse? Why? (1 points)
  + Use the sum of two targets two predict cnt. Estimate the quality. Is it higher than for one target model? (1 point)
* Build regression model with cnt as target and estimate the quality on last half of the year / month.
* In the process of completion, answer the following questions:
  + How the temperature and humidity influences rental count rates? What is more important - temperature or feeling temperature? (2 point)
  + What are the most influentual features? (2 points)
  + What is more important: current season or current hour? (1 point)

Total points: 20. Threshold for completion: 15 points.