**DATA PREPARATION**

The specific ML algorithm you are working (Linear regression in our case) with and the knowledge of the problem domain will influence this step and you will very likely have to revisit different transformations of your preprocessed data as you work on your problem.

Data preparation is a large subject that can involve a lot of iterations, exploration and analysis. Getting good at data preparation will make you a master at machine learning.

**N.B.** For data cleaning task we will use Python libraries: Pandas & Numpy (so please include it in your Python scripts)

**DATA PREPARATION TASK**

1.      Create Python script for data preparation:

-        read data from BASE/<YYYY>/MM/DD

- Concatenate partial data to one data frame.

- Deal with Scaling problem

**Scaling**: The preprocessed data may contain attributes with a mixtures of scales for various quantities such as dollars, kilograms and sales volume. Many machine learning methods like data attributes to have the same scale such as between 0 and 1 for the smallest and largest value for a given feature. Consider any feature scaling you may need to perform.

* Scale PoolsQC column to float values, for example:

“ex” – 1.0, “gd” – 0.8, “ta” – 0.5, “fa” – 0.2, “po” – 0.0

* Scale PavedDrive column to:

“y” – 1.0, “p” – 0.5, “n” – 0.0

- Deal with Decomposition problem

**Decomposition**: There may be features that represent a complex concept that may be more useful to a machine learning method when split into the constituent parts. An example is a date that may have day and time components that in turn could be split out further. Perhaps only the hour of day is relevant **to the problem being solved. consider what feature decompositions you can perform.**

* Decompose DateSold to month and year components in separate columns.
* Remove DateSold column.

- Deal with Aggregation problem

**Aggregation:** There may be features that can be aggregated into a single feature that would be more meaningful to the problem you are trying to solve. For example, there may be a data instances for each time a customer logged into a system that could be aggregated into a count for the number of logins allowing the additional instances to be discarded. Consider what type of feature aggregations could perform.

* Aggregate total basement surface:

Sum of columns “BsmtFinSF1”, “BsmtFinSF2”, “BsmtUnfSF”

* Aggregate total living area:

Sum of columns “1stFlrSF”, “2ndFlrSF”, “LowQualFinSF”

* Aggregate age of house:

Subtract “YearBuilt” from this year

- save data to “PREPARED/<YYYY>/

2.      Scheduling/automatization of data preparation script

- Create Scheduling task with Cron job (to collect data daily)