**Dataset summary:**

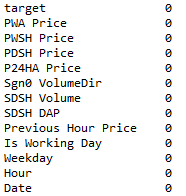
Dataset includes 13 features in total with 1 target variable (Electricity price per hour). The goal is to predict the electricity price per hour using the predictor variables.

**Data Pre-processing:**

**(i) Test for missing data**

The electricity dataset is tested for any missing values. It was found that there were no missing values.

**Information on missing value**



**(ii) Correlation**

We perform test of Pearson correlation to identify if there exist multicollinearity among the predictor variables. In regression models, it is important to test for existence of multicollinearity. If there is high correlation among a pair of predictor variables, then multicollinearity exists among them.

**Pearson Correlation**



It was found that high amount of correlation exists between the following pair of predictors

- SDSH DAP and Previous hour price (78%)

- PWA price and P24HA Price (68%)

- P24HA Price and Previous hour price (66%)

So, yes multicollinearity exists among the variables and how it can be handled?

(iii) **Handle multicollinearity using PCA**

Principal component analysis (PCA) is often used to handle multicollinearity among the variables. In this process, the dimension of the dataset is reduced to contain only the significant factors. It gives a optimally weighted combination of a group of correlated variables.

The explained variance of our first 4 PCA components were: 0.57, 0.21, 0.08, 0.06 and this sums up to 92% of significant information. The remaining 7 of the components can be removed from the machine learning that we are going to build.

**(iv) Split the dataset**

This dimension reduced dataset is then split with 80% for training and 20% for testing. The model will be first trained and then used for prediction (testing purpose).

**(v) Model building**

In our model, we use Decision tree and k-nearest neighbours (KNN) for training and prediction. Ours is an ensemble model, so we combine the results of both the algorithms using the Ensemble model averaging method.

**(vi) Model evaluation**

We evaluate the model performance using the standard metric for regression problems i.e., Root mean squared error. The RMSE value for the training set is 743.24.

**(vii) Target value prediction of test set**

We have trained the model in the earlier stage and now we will use this trained model to make prediction for the test set. We cannot evaluate this model using RMSE because we do not have the actual values. We can only compare this performance with other standard algorithms like SVR, & Multiple linear regression. Excerpt from final prediction is below:

