Assignment #2

The data scientists at KMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined.

Kmart wants to better understand the properties of products and stores which play a key role in increasing sales.

**Goal**: The aim is to build a predictive model to find out the sales of each product at a particular store.

**Answer**:

1. What steps should be taken for pre-processing the data?

Data set have categorical as well as numerical data and missing values also so I have followed the below steps

* First pre-processed ‘Outlet\_Size’ feature and find out the missing values by ‘counts’ and ‘freqs’ of categories and then filled missing values accordingly.
* Feature ‘Item\_Weight’: Find the total ‘NA’ value and then fillna value the mean value
* Feature ‘Item\_Fat\_Content’: Found Irregular values, used replace method to regularize it.
* Feature ‘Item\_Visibility’: Found total zeros and replace zeros with the mean values of Item\_Visibility.
* Features ’Item\_Fat\_Content’, ‘Outlet\_Size’, ‘Outlet\_Type’ ‘Outlet\_Location\_Type’: Used map method to label encoding
* Feature ‘Item\_Type’ , ‘Outlet\_Identifier’: get dummies values by pd.get\_dummies
* Feautes 'Outlet\_Identifier', 'Item\_Outlet\_Sales', ’Item\_Type’: dropped
* Same above steps done for test data set

1. Explain your modeling approach:
   1. Choice of ML algorithm

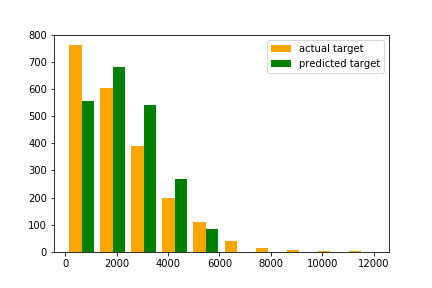
* As the target variable is known so it is surely a supervised problem and target variable is continuous so algorithm will be from regression.
  1. Feature Selection
     + Generated the hypothesis testing and then verified those testing with actual data
     + Store level hypothesis: city type, population density, store capacity, competitors, marketing, location, ambience
     + Product level hypothesis: branch, packaging, utility, Display area, advertising, promotional offers, visibility of store.
     + There were 16 hypothesis testing where 9 are present in actual data
     + Find the correlation of independent variables.

1. How would you evaluate the performance of your model? Explain the choice of metrics used.

* Evaluated the performance of model by R squared score and mean squared error
* R squared score tells about the fitness of model
* Mean squared error tells the errors in the model
* Fit training data set into six regression model and achieved 65% score with RandomForestRegression
* Algorithm used to train the model Linear regression, Lasso, Ridge, ElasticNet, Random Forest Regression, Sequential(Keras)

1. Show plots for visualizing your model performance.

* Plot the histogram graph to check the performance of a model

****