**Retail Price Optimisation**

**Problem Statement:**

Managing prices by the retail industry is the key to profit in business. It is essential to identify a reasonable price range and make adjustments for the prices of products accordingly.

The user need to build best model takes in sales data, product Characteristics, on the structure data, textual information, and pricing rules and several inputs. The model suggest price for a product in real-time by considering thousands of relationships within a product.

**Used Dataset:**

Model.csv which is related to the prices and properties of products.

Dataset taken from kaggle.

**Steps Involved:**

1.Importing dataset

2. Data Pre-processing

3. Data Visualization

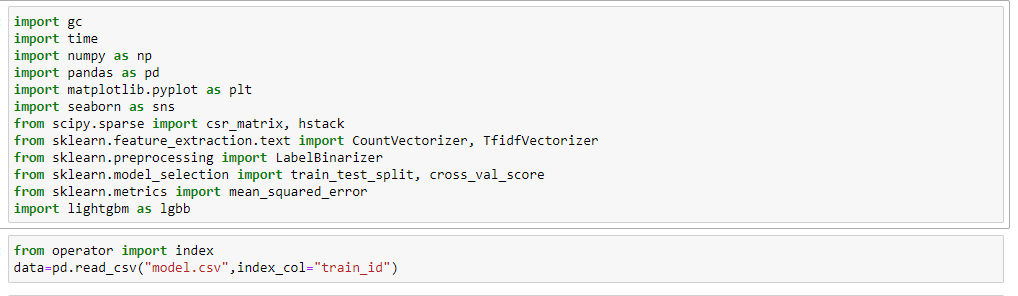
4. Model building

5. Testing the model

**Process:**

**Importing dataset:**

Firstly imported all the necessary modules. Loaded model.csv file to DataFrame named ***data***.

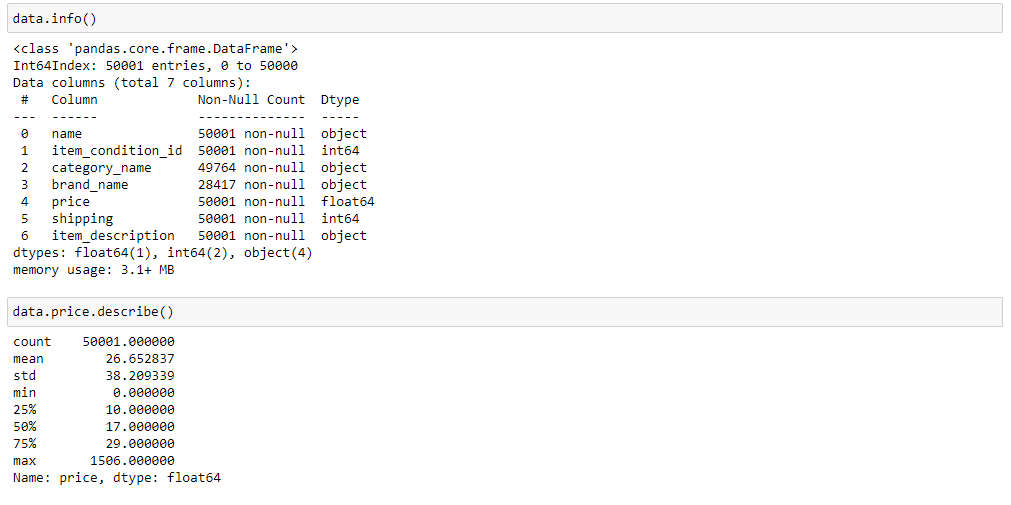


->observed all the data attributes and their data types.

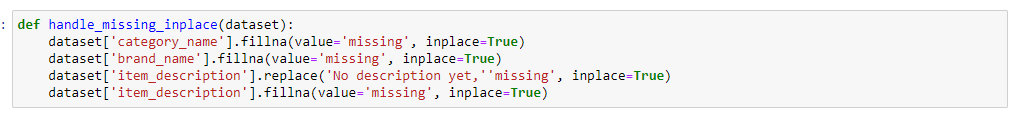
There are 7 columns namely name, item\_condition\_id , category name , brand name , price , item\_description , shipping .

1 float datatype, 2 integer datatypes, 4 object datatypes are found

->observed dependent variables using describe function.

**Data Pre-processing:**

->In the data we found that there are some missing values in category\_name , brand\_name , item\_description. Filled missing values with the value “missing” and in the column item\_description there are both missing and No description yet values filled both of them with the values “missing”.



-> Converted all the above columns into category datatype.

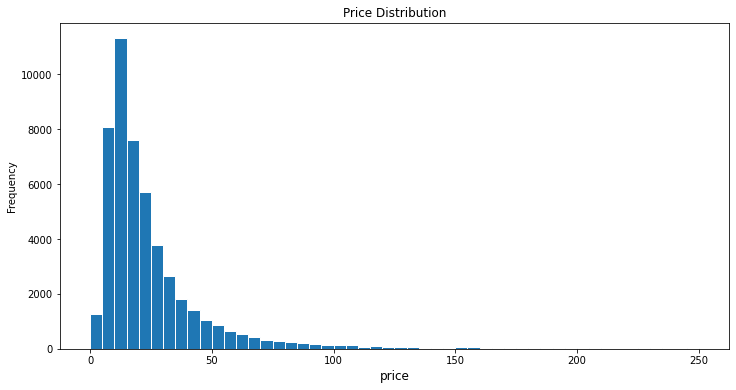


-> Removed all the rows having price 0.

**data= data[data.price != 0].reset\_index(drop=True)**

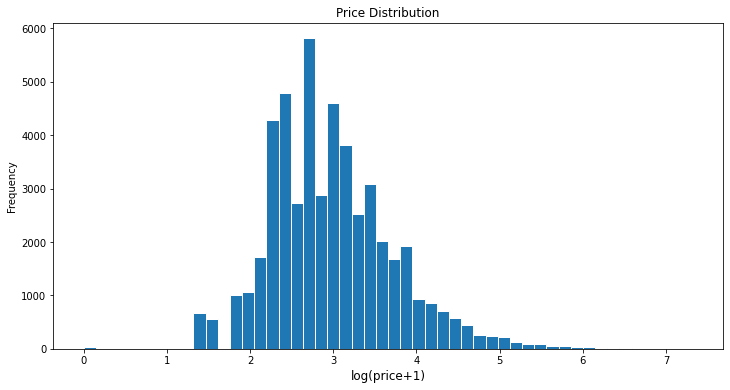
**Data visualization:**

Histogram of price:

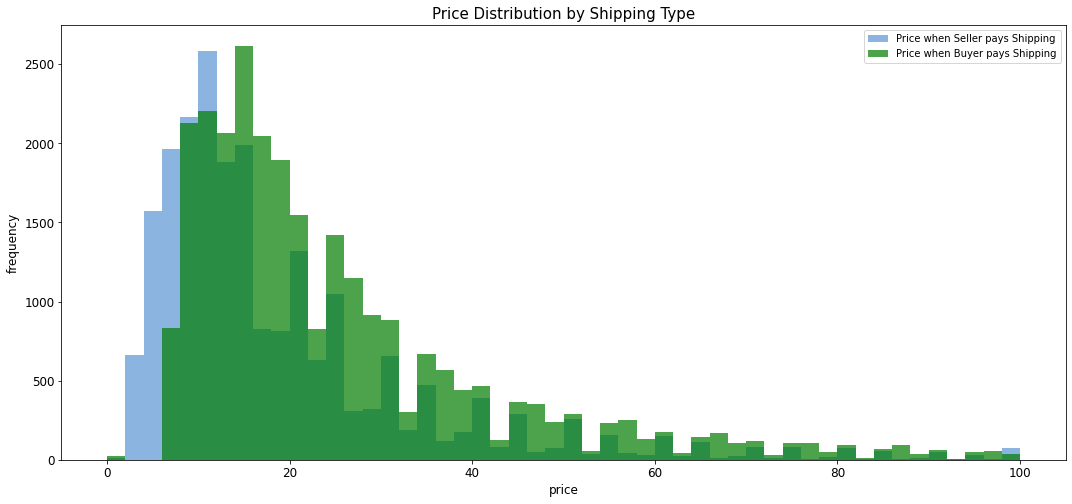


Here we found that price is right skewed data so we will apply log transformation on it.

Price after applying Log transformation:



Price distribution with respect to shipping type:

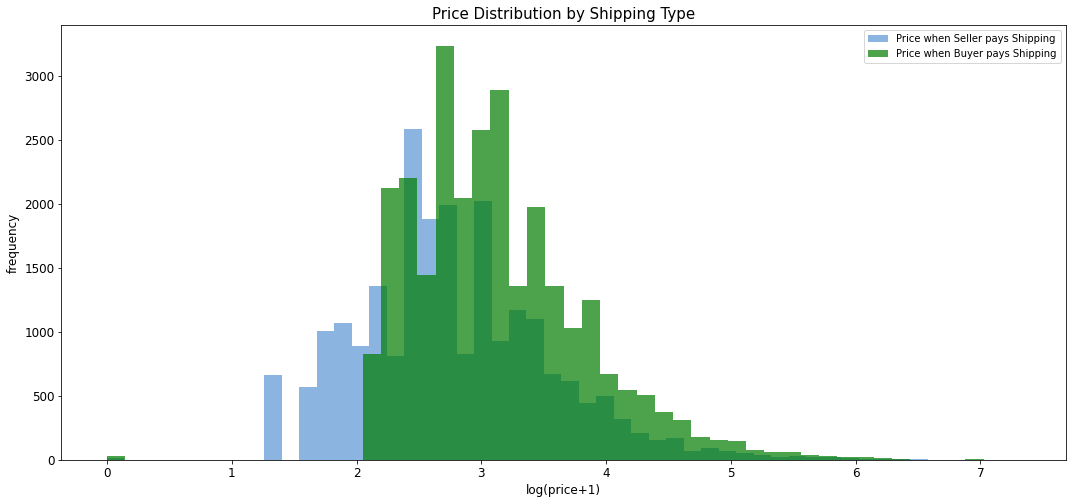


From the histogram we found that

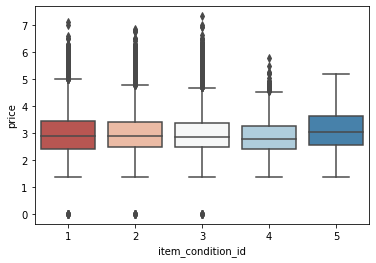
The average price is 22.42 if seller pays shipping

The average price is 30.09 if buyer pays shipping

The price is right skewed in this case too. So apply log transformation.

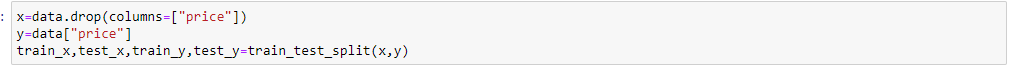


Item condition id with respect to price:

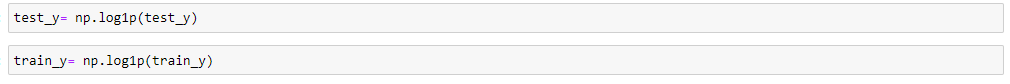


**Model Building:**

->Splitted the data into train and test data in the proportion of 75:25.



->Applied log function on price column to make it normalized.



-> Applied countvector() on category name to convert a collection of text documents to a vector of term/token counts.

->Applied tdifvectorizer() on item\_description to map the most frequent words to feature indices and hence compute a word occurrence frequency (sparse) matrix.

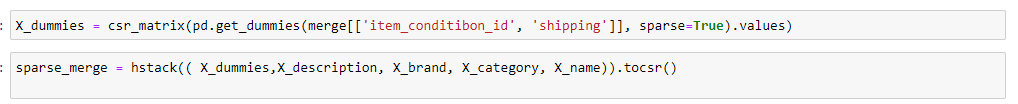
->Applied labelbinarizer() on brand\_name to accepts Categorical data as input and returns an Numpy array.

->Applied getdummies() on item\_condition\_id and shipping then converted it into csr matrix.





->stacked sequence of X\_dummies,X\_description, X\_brand, X\_category, X\_name horizontally using hstack() and converted it into csr matrix.

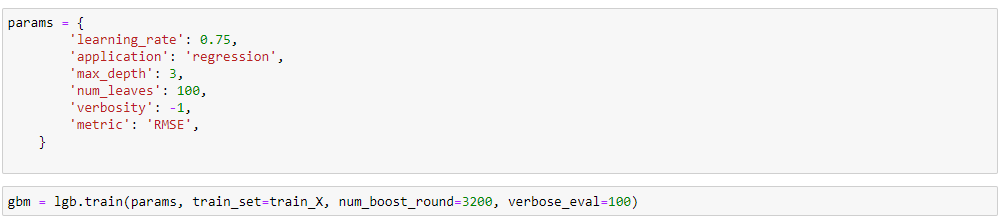


->Removed features with document frequency less than one.



->Used Lightgbm algorithm for the prediction.

Trained it using the 75% proportion data.



**Testing Model:**

-> Validated the model with remaining 25% data and got RMSE value of 0.582625956002516.

