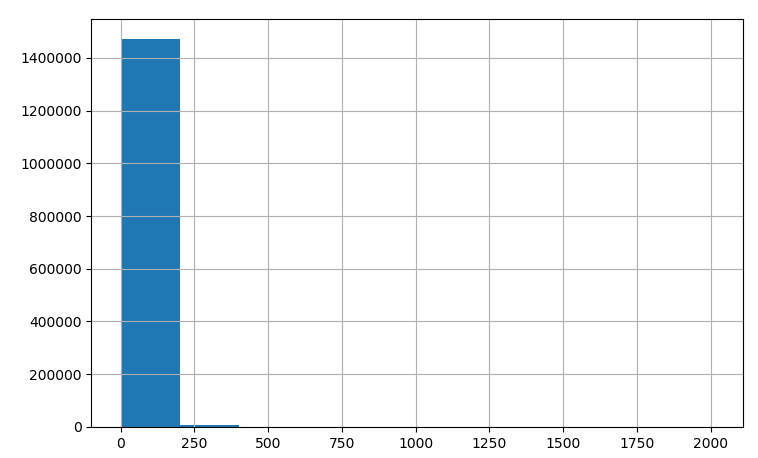
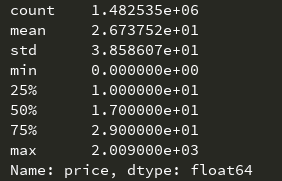
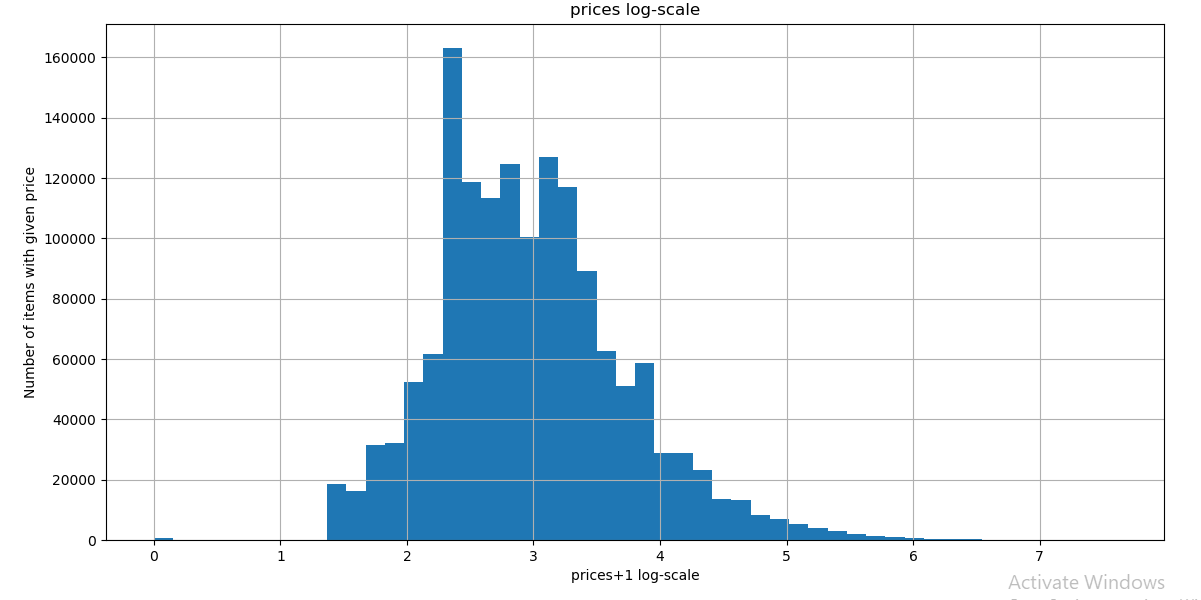
Presenting Information about the Data:

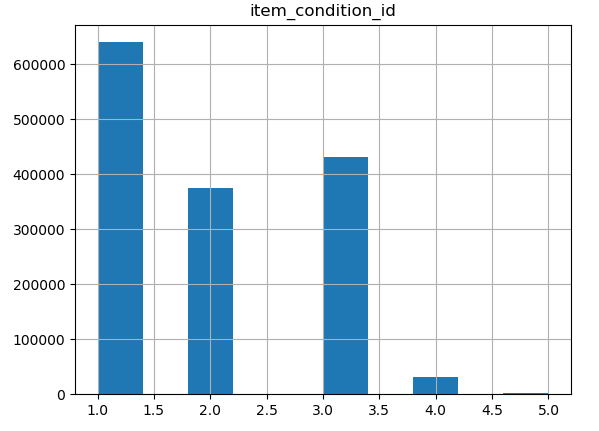


* It can be seen that Most prices of the items are ranged from 1$ - 300$, and another small fractions moves between 200$-450$ (approx.)

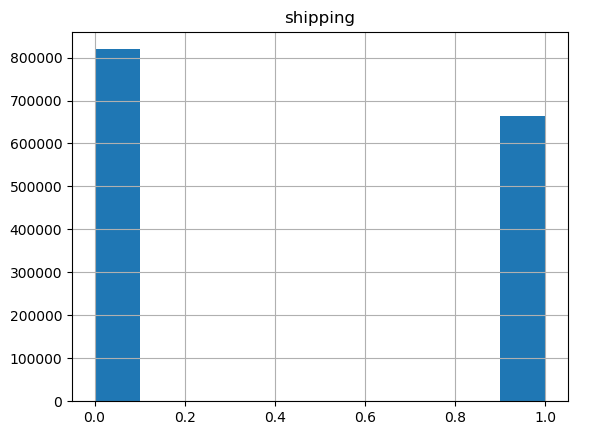
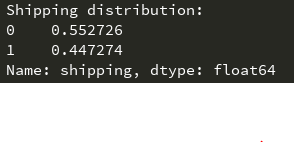
This information could be useful to assist with evaluating a good\decent\really-bad model using their MSE\ MAE.



Presenting the log-scale of the prices values , can show us the distribution of the prices in a better way. (added +1 to each price to prevent zero values)



* We can see that almost half of the data rows contain item\_condition=1. The other half have it’s value on 2 / 3. Kind of a weak distribution. May not be too helpful in our opinion in the prediction.

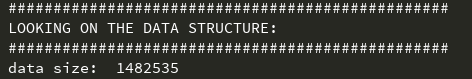


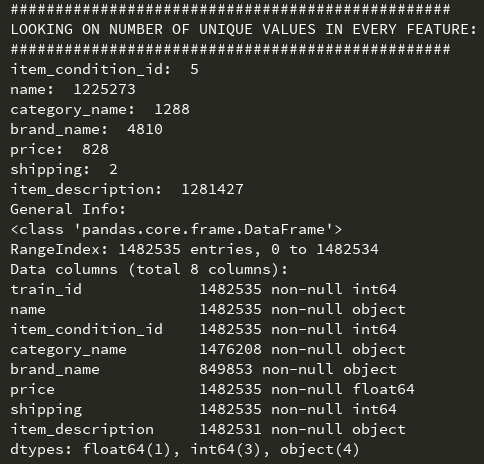
The ‘shipping’ feature is almost evenly distributed.

* Rest of the features are text, and therefore will be converted into vectors using some kind of sentence embeddings (Eventually used InferSent)

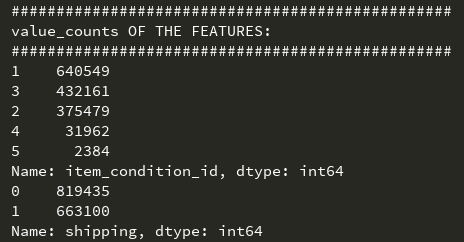
Some more general information:

The information is being calculated in ‘present\_data.py’:

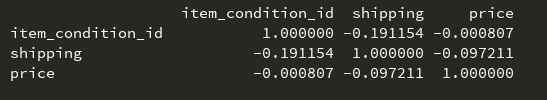




Categoric features different values counting :



Correlation Matrix of the initial numeric features:



We can see that shipping has a certain correlation with the price. Close to 10%.

The condition of the item (ranging from 1 to 5) not too much correlation by itself.

Checking each one of the other features correlation to the price may be possible with a linear regression model using the infersent encoding of the feature. Conclusions of this will be put here after all of the data encoding works (encoding too much data doesn’t work on our computers).