Steps

1. Data analysis
2. Cleaning up the data
3. Features extraction
4. Identify ML model
5. Train and validate the model
6. Tuning the model and feature extraction
7. Use the model to predict the price for test.csv
8. Data analysis/cleaning up the data
9. Numerical features

* Price – This is our response that we need to predict
* Shipping cost

1. Categorical features

* Shipping cost: A binary indicator, 1 if shipping fee is paid by the seller and 0 if paid by buyer.
* item\_condition\_id: The condition of the items provided by the seller
* name: The item's name
* brand\_name: The item's producer brand name
* category\_name: The item's single or multiple categories that are separated by "\"
* item\_description: A short description on the item that may include removed words, flagged by [rm]

**Price(Target Variable)**

The median price of all the items in the training is about \$267 but given the existence of some extreme values of over \$100 and the maximum at \$2,009, the distribution of the variables is heavily skewed to the left. So let's make log-transformation on the price (we added +1 to the value before the transformation to avoid zero and negative values).

**Shipping**

The shipping cost burden is decently split between sellers and the sellers (55%) pay buyers with more than half of the items’ shipping fees.

**Item Category**

There are about 1,287 unique categories but among each of them, we will always see a main/general category firstly, followed by two more particular subcategories (e.g. Beauty/Makeup/Face or Lips). In addition, about 6,327 items do not have a category labels.

**Brand name**

There are 4809 unique brand names in the training dataset.

**Item Description**

It will be more challenging to parse through this particular item since it's unstructured data. Does it mean a more detailed and lengthy description will result in a higher bidding price? We will strip out all punctuations, remove some english stop words (i.e. redundant words such as "a", "the", etc.) and any other words with a length less than 3:

We also need to check if there are any missing values in the item description (4 observations don't have a description) andl remove those observations from our training set.

If we look at the most common words by category, we could also see that, size, free and shipping is very commonly used by the sellers, probably with the intention to attract customers, which is contradictory to what we have shown previously that there is little correlation between the two variables price and shipping (or shipping fees do not account for a differentiation in prices). Brand names also played quite an important role - it's one of the most popular in all four categories.