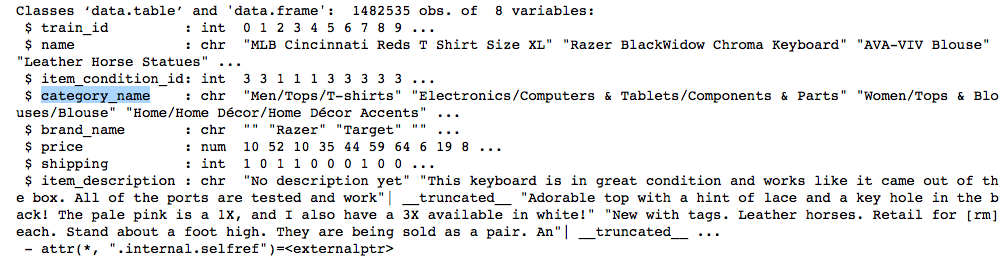
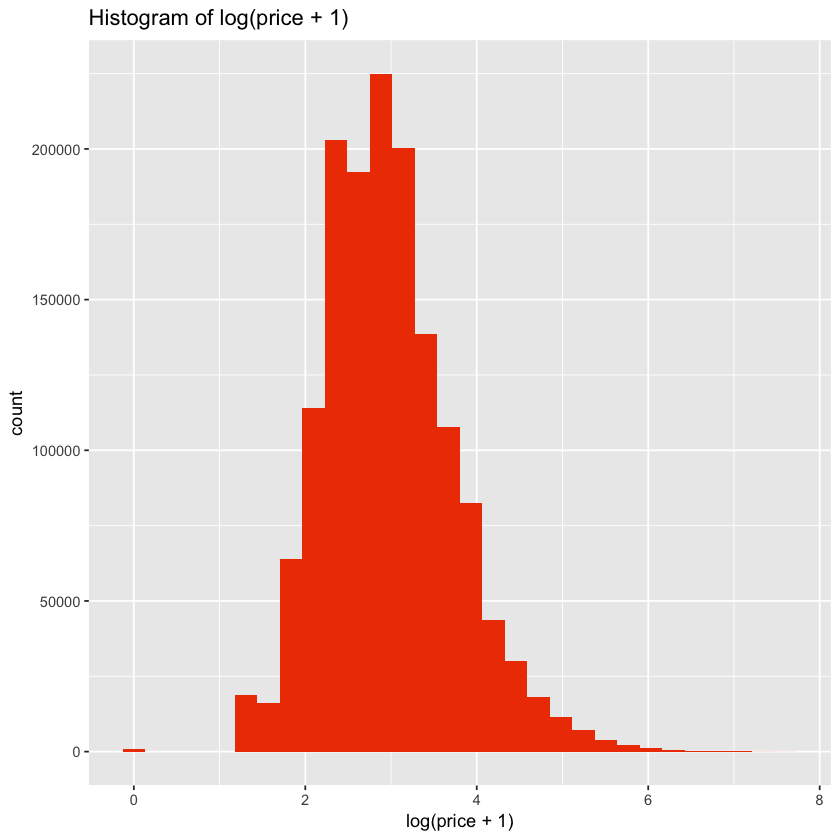
EDA

1. General

One thing that we supposed to do before we go deep into EDA is having a general look of the data set.

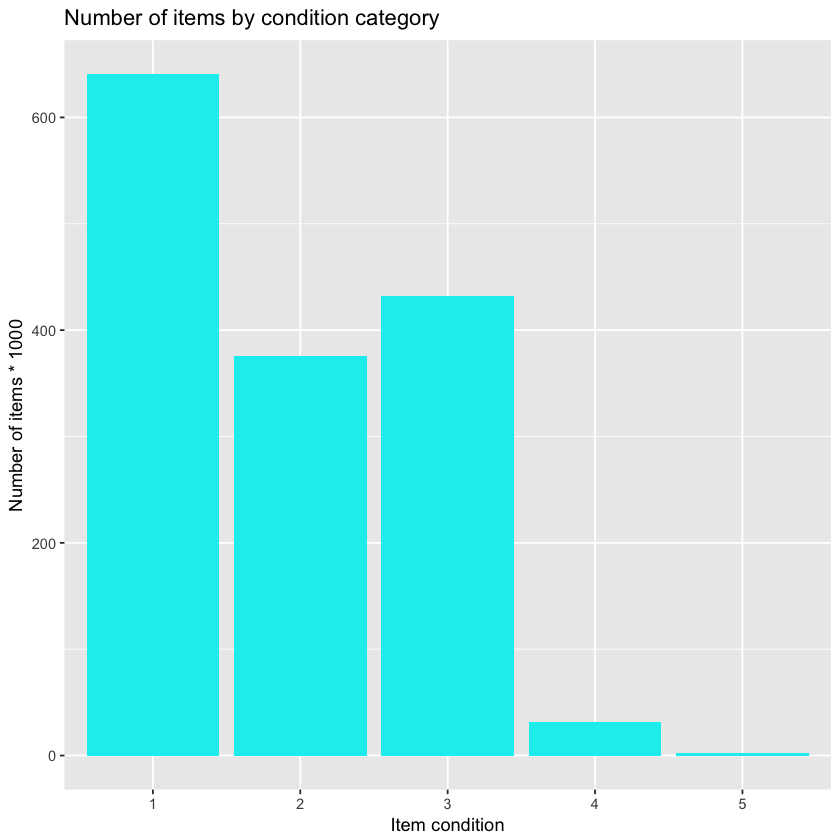


By look through all the eight columns, there are two of them draw my attention tightly. First is the ‘item\_description’ which need NLP to handle with. Another is the ‘category\_name’ which mostly consists of three levels, seems need some tricky strategies.

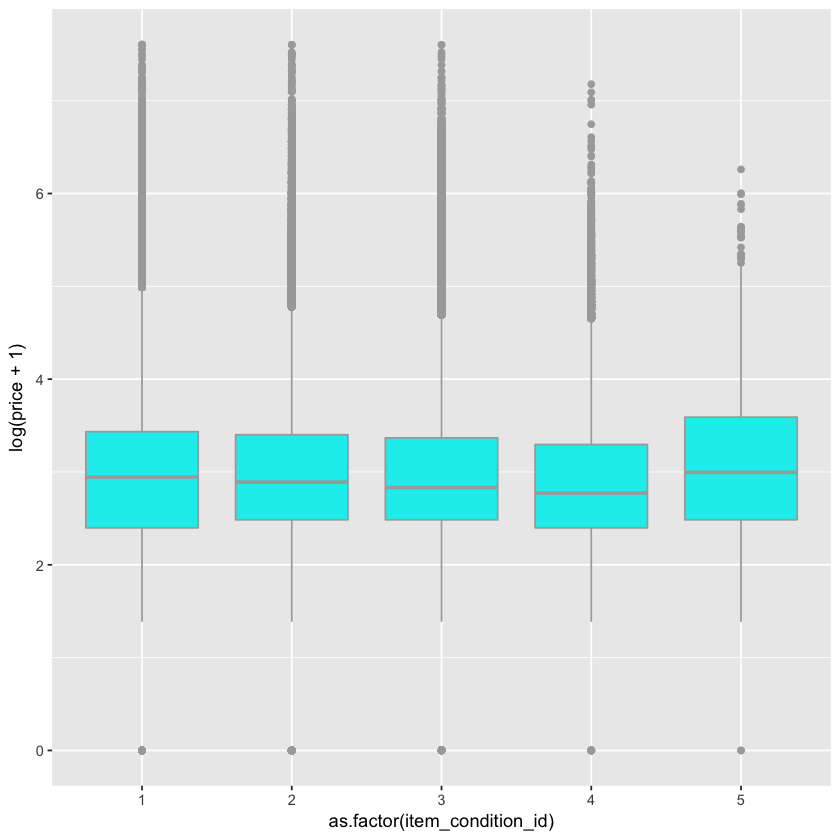
2. Price

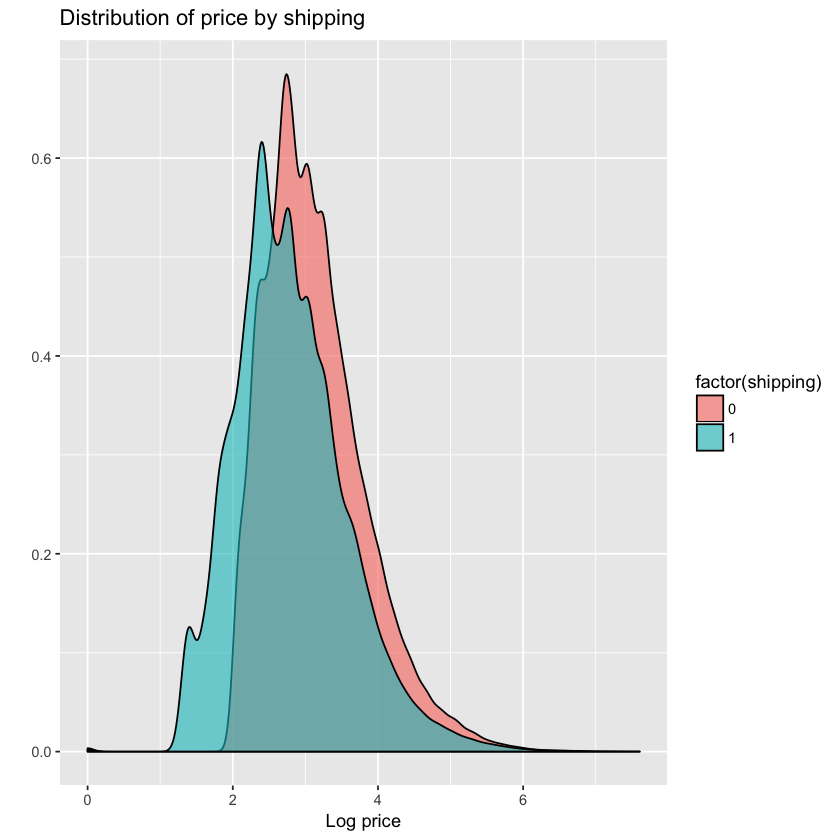
Primarily, let’s have a look of our target—price. The price ranges from 0 to 2009. Because price is likely skewed and because there are some 0s, we’ll plot the log(price+1).

As the image shows that over 80% is between 2 to 4.

3. Item Condition

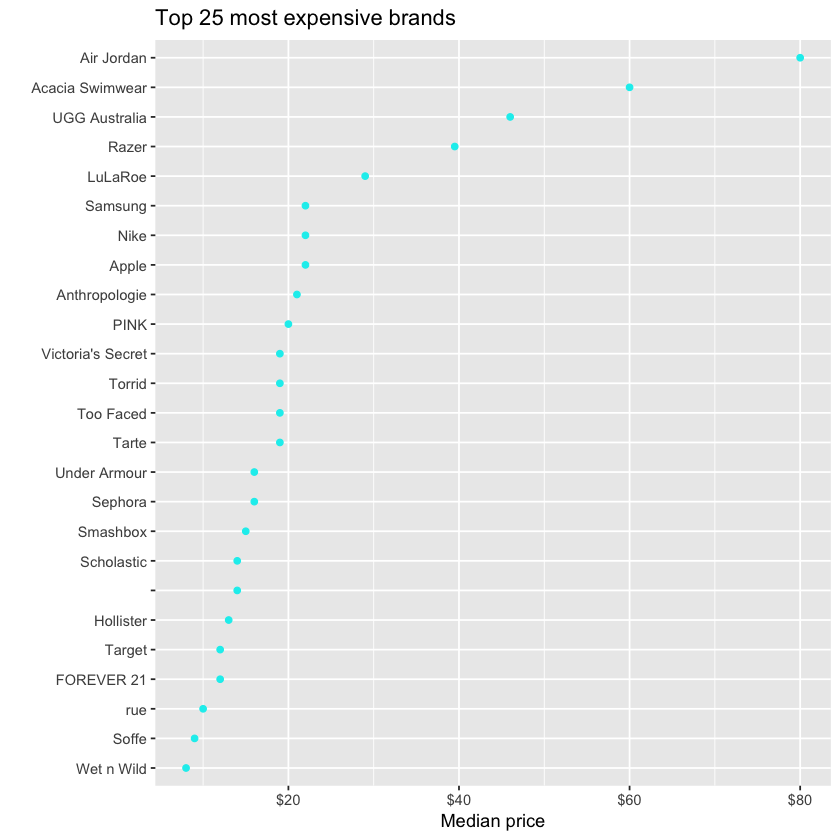
The item condition ranges from 1 to 5. There are more items of condition 1 than any other. Items of condition 4 and 5 are relatively rare. It’s not clear from the data description what the ordinarily of this variable is. My assumption is that since conditions 4 and 5 are so rare these are likely the better condition items. We can try and verify this. If a higher item condition is better, it should have a positive correlation with price. Let’s see if that is the case.

Looking at the average price by condition shows a relationship that is not quite as neat as I had hoped. Condition 5 clearly has the highest price, however condition 1 has the next-highest price, followed by condition 2, then 3, then 4. Consider with the numbers of items for each condition, we can easily know that the condition is decreasing through 1 to 5. The reason that the price of condition 5 is higher than other’s is the uncertainty cause by the data insufficient.

4. Shipping info

My initial thought is that items where the shipping fee is paid by the seller will be higher-priced. However, there are numbers of conflating factors. This may be true within specific product categories and item conditions, but not when comparing items on the aggregate. Let’s see.

From the plot, we can ensure that the price will be higher if shipment fee is paid by seller.

5. Brand

The Air Jordan and Acacia Swimear brands are by far the most expensive brands, with a median price of $80 and $60 respectively.

Nearly half of the items don't have brands. The proportions of items that have brands vary in different categories. For example, nearly all handmade items don't have brand names, of course.

For brands, they are not in a hierarchical order and there are too many to be fitted in one graph. So, I plotted the count of top 10 most frequent brands for a rough look. Each brand contains items from 1 or more major categories. Not surprisingly, the top brands are dominated by women items except Apple and Nintendo.

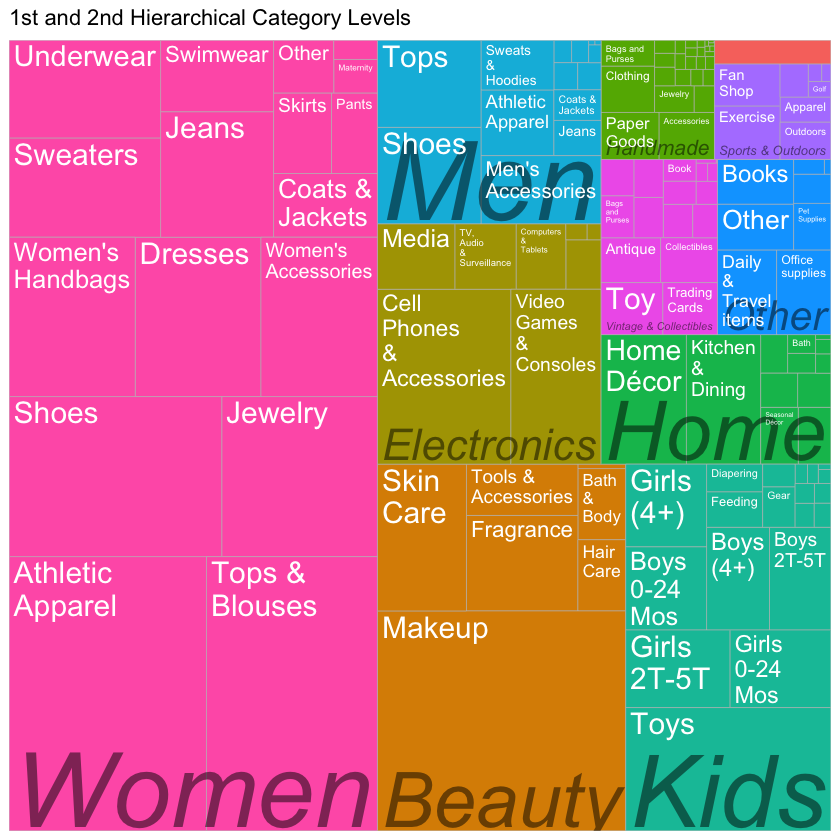
6. Category name

We notice that the category\_name is actually encoded as three or four hierarchical levels splitted by /. (Thanks to Abhinav Reddy Kaitha there are some items with four levels instead of three)

We can split the category names and store them into 4 columns. The major category (1st category) only has 11 levels and we can make distinguishable visualizations on them. From the 2nd level on the # of levels are too many to visualize.



Most items only have three levels of categories. But the 4th level exists with 8 unique sub-categories and 4389 items. For modeling perspective, it may be fine to combine it with 3rd levels but for analysis purpose I extract and keep the 4th level here.

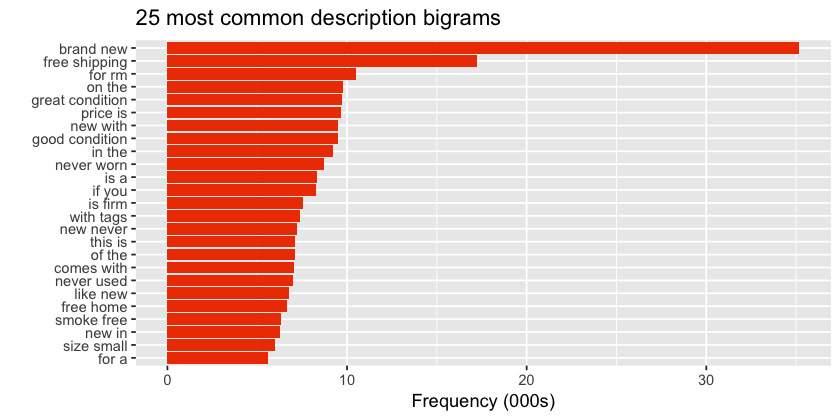
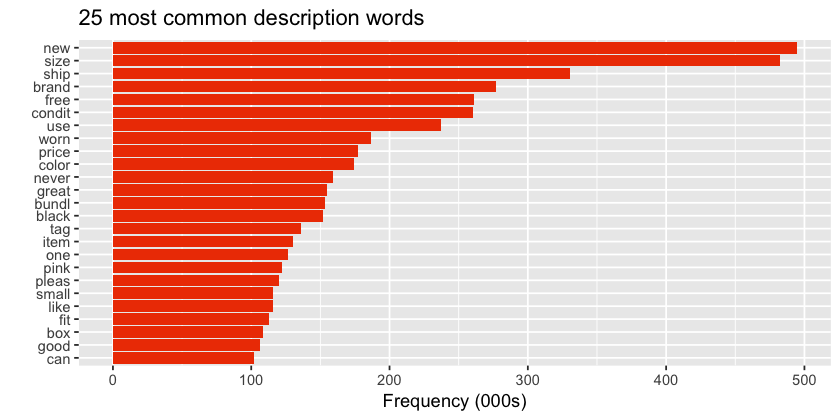


For visualization, we plot the 1st and the 2nd level of categories to show the distribution of the number of the retail items. We can see that the majority, nearly a half, items belong to Women category. So, we than plot the 2nd and 3rd layer of the Women category to have a deeper look.

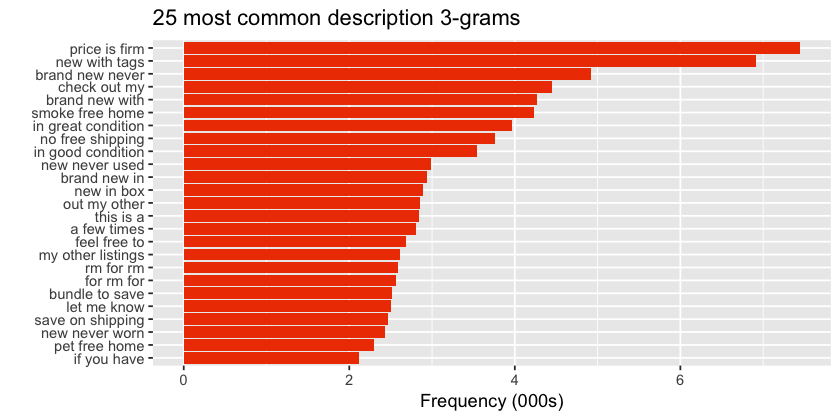
7. Item description

For item description, we have mainly three steps: remove the price description; remove English stop words, punctuation, and stem words; using N-grams to find out most common grams when N=1,2,3.

N = 1 N = 2



N = 3



TF-IDF

TF-IDF stands for term frequency-inverse document frequency, indicates the importance of a word to a document in a corpus. Similar with the weight factor when we select features in Neural Network and Random Forest, TF-IDF can help us filter the words that only appear few times and the words appear frequently but have no significant meaning.