Mercari price prediction

Vinh Vu

I. Introduction:

Product pricing at scale is a difficult task, considering just how many products are sold online. Clothing has strong seasonal pricing trends and is heavily influenced by brand names, while electronics have fluctuating prices based on product specs. Through this project, I want to build an algorithm which can suggests product prices for online retailers.

My client, Mercari, is one of Japan's biggest online shopping sites. In order to improve better serve their customers, Mercari wants to offer price suggestion to their sellers when they are posting a new product for sale on their site. Based on my prediction, Mercari would be able to:

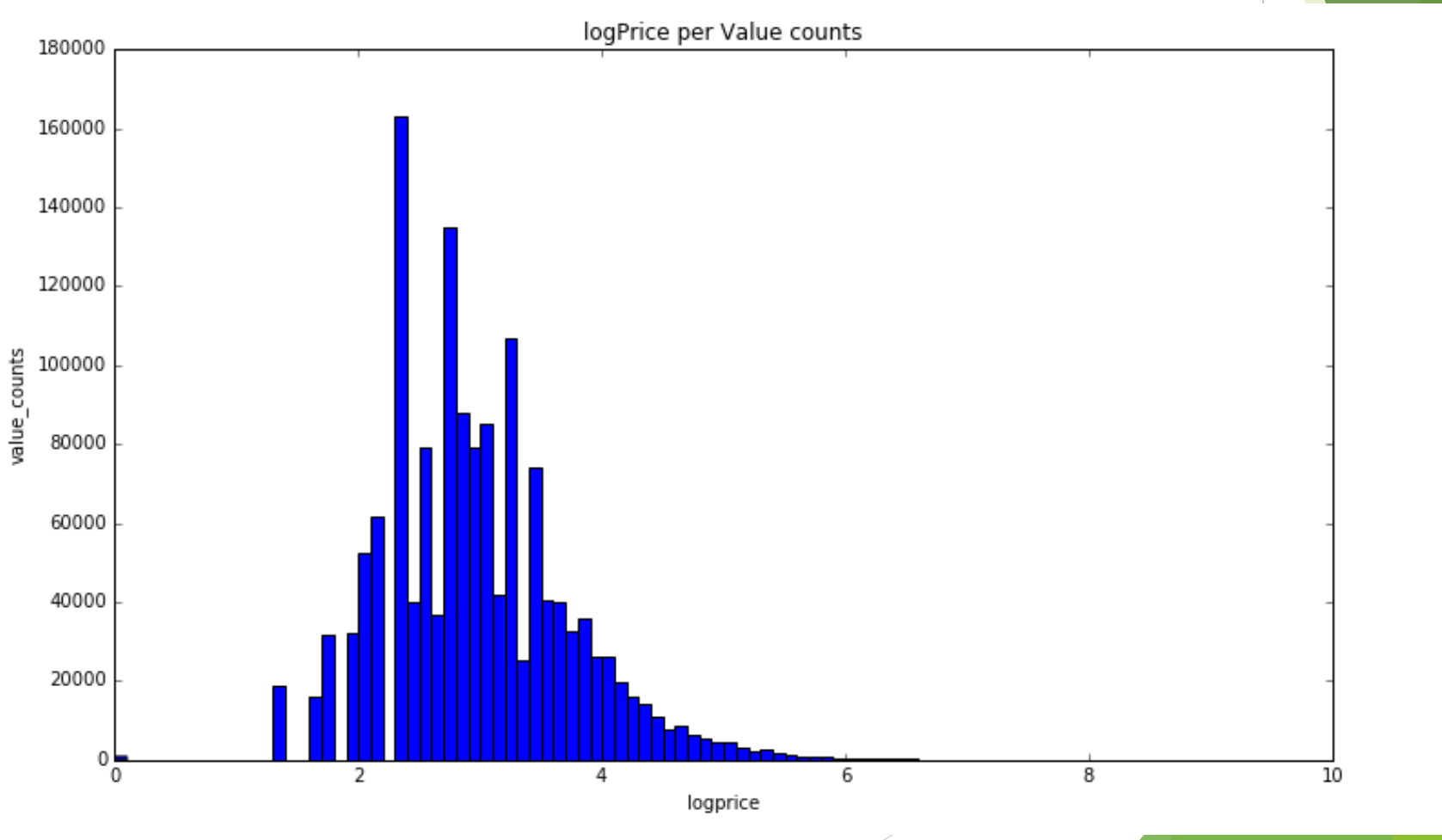
- Identify the product detail such as category, brand and type.
- Offer a price suggestion based on products' detail.
- Modify the product price before selling on the Mercari marketplace.

II.Data obtaining and cleaning

- Mercari dataset which is provided by Kaggle, include train.csv, and test.csv. There are some missing values in the dataset as well and I imputed them with Nan's for simplicity. Here are some of the pre-processing steps I did:
- Handling Missing Values – Replaced with Nan.
- Lemmatization performed on item description.
- Label Encoding – Turned categorical columns into 0's and 1's.
- Tokenization – Given a character sequence, tokenization is the task of chopping it up into pieces.
- Scaling – Scaled the price variable (log price).

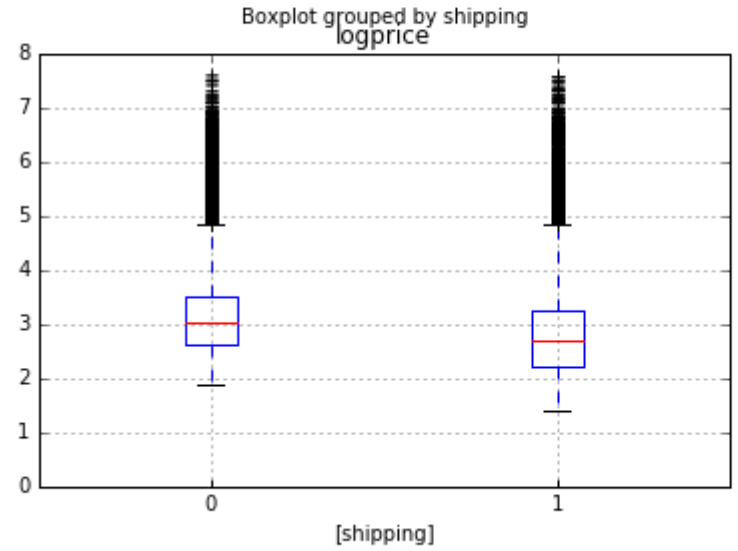
III.Data Story:

1.Price:

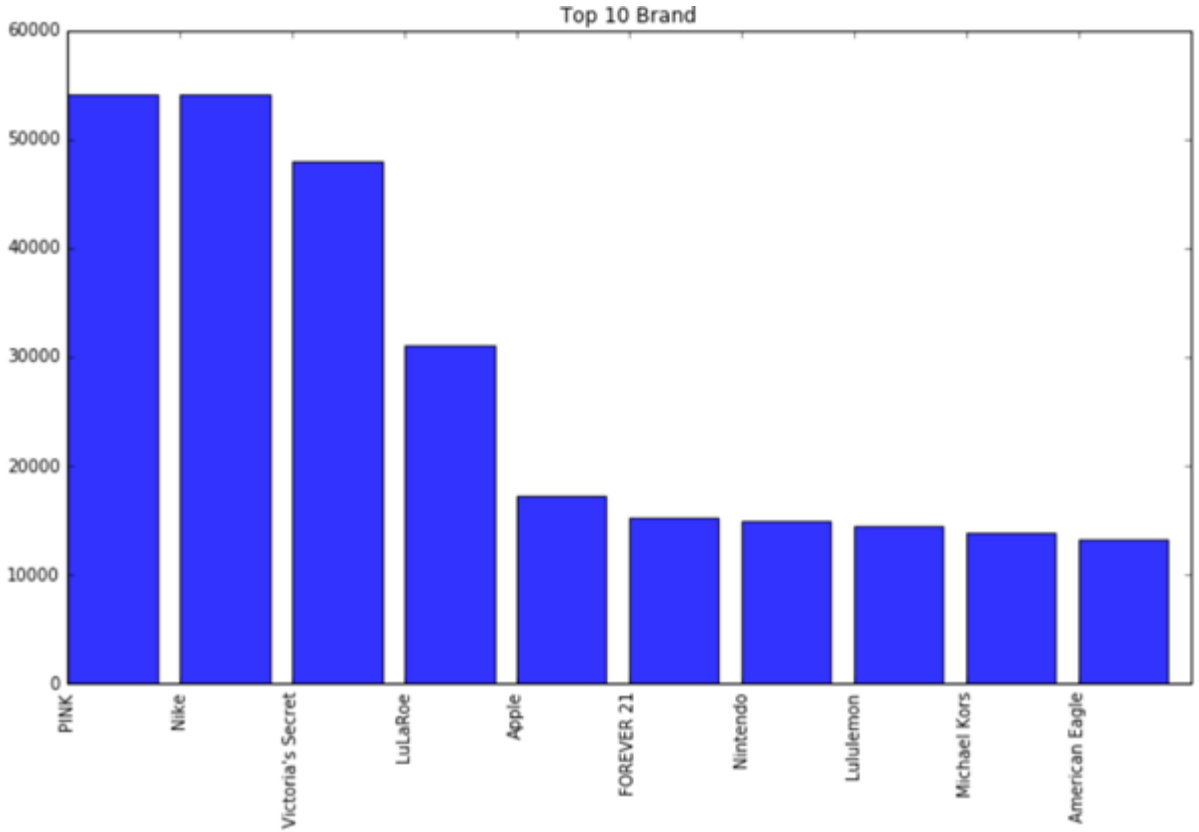


2.Log Price with shipping:

	mean	std
shipping		
0	3.133894	0.693802
1	2.787720	0.770638



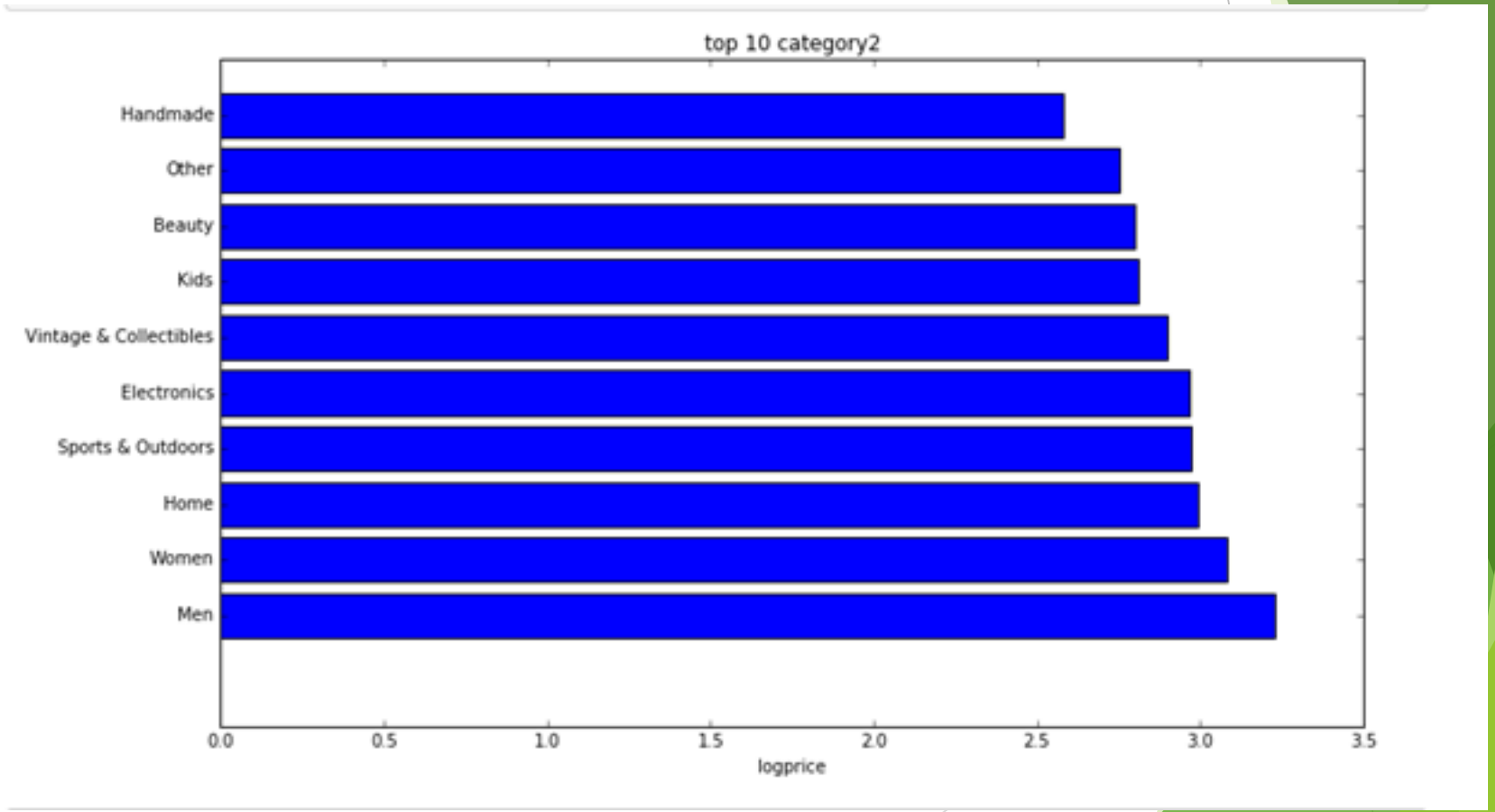
3. Top Brand Distribution:



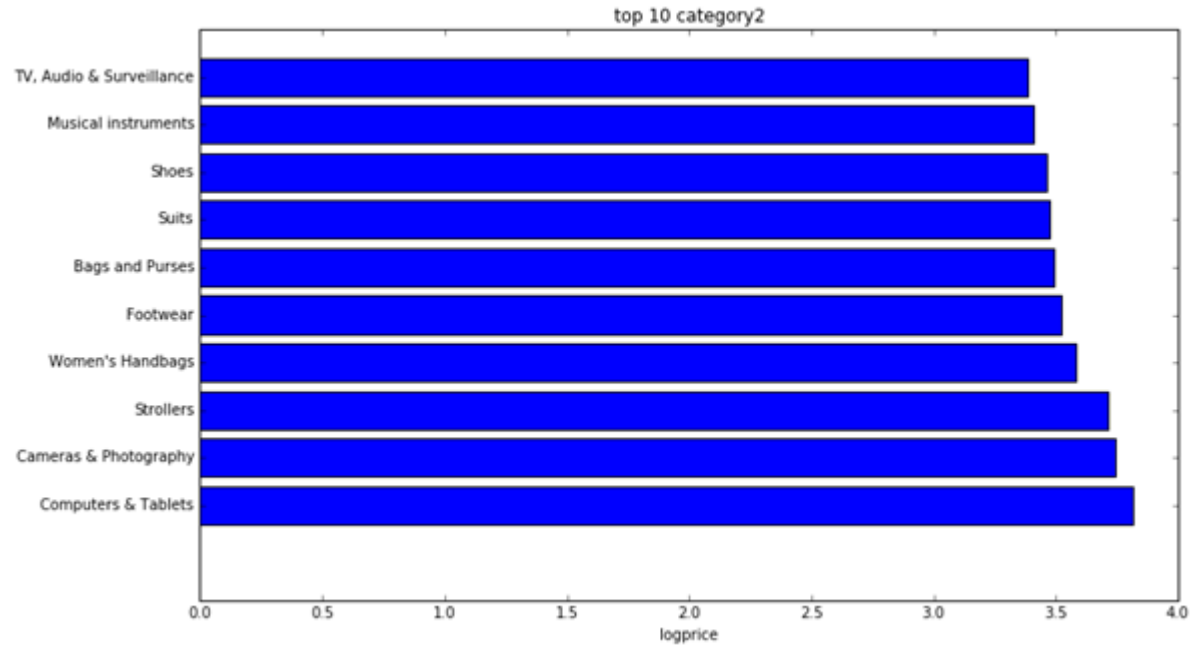
- Pink
- Nike
- Victoria’s secret

4.Category:

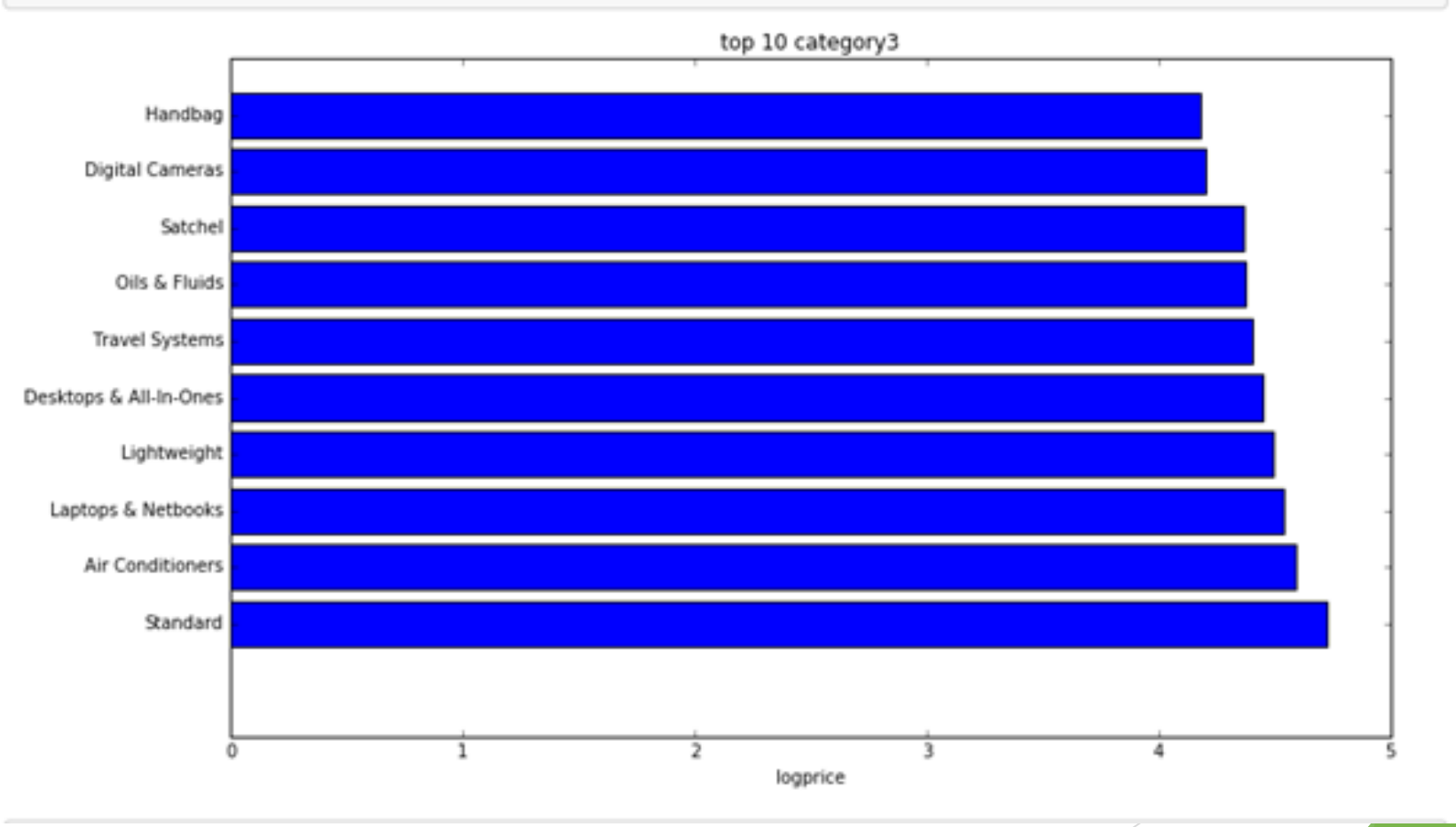
Category 1:



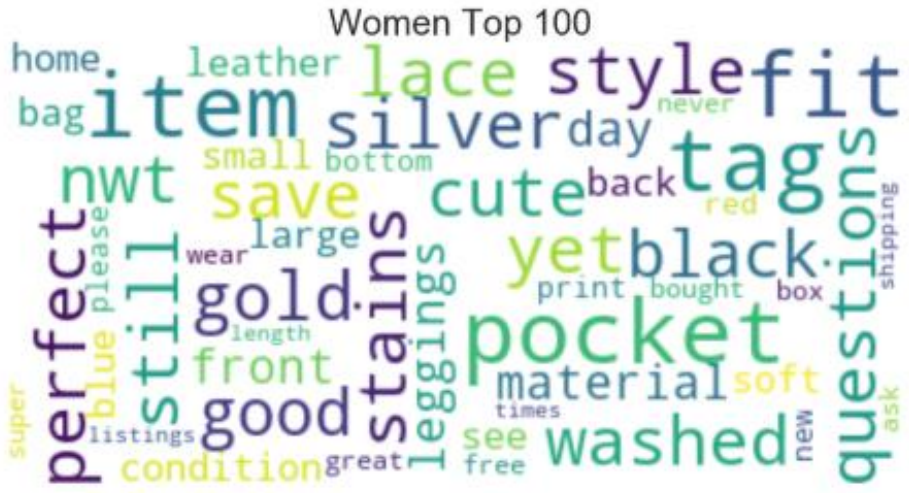
Category 2:

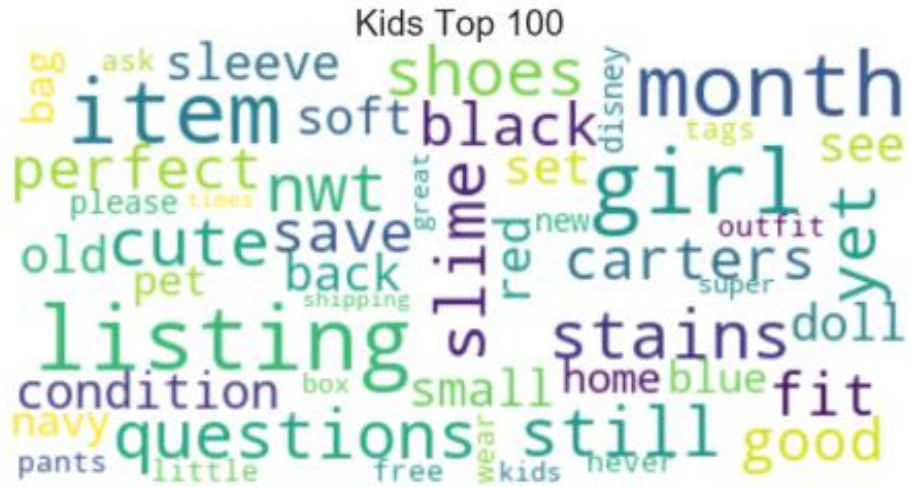

top 10 category2
TV, Audio & Surveillance
Musical instruments
Shoes
Suits
Bags and Purses
Footwear
Women's Handbags
Strollers
Cameras & Photography
Computers & Tablets
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
logprice

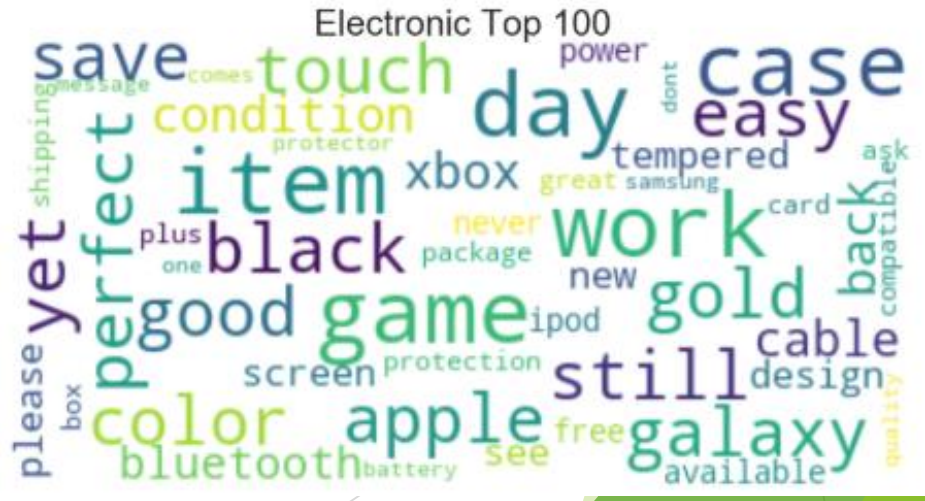
Category 3:


top 10 category3
Handbag
Digital Cameras
Satchel
Oils & Fluids
Travel Systems
Desktops & All-In-Ones
Lightweight
Laptops & Netbooks
Air Conditioners
Standard
0
1
2
3
4
5
logprice

5.Item Description by Category:

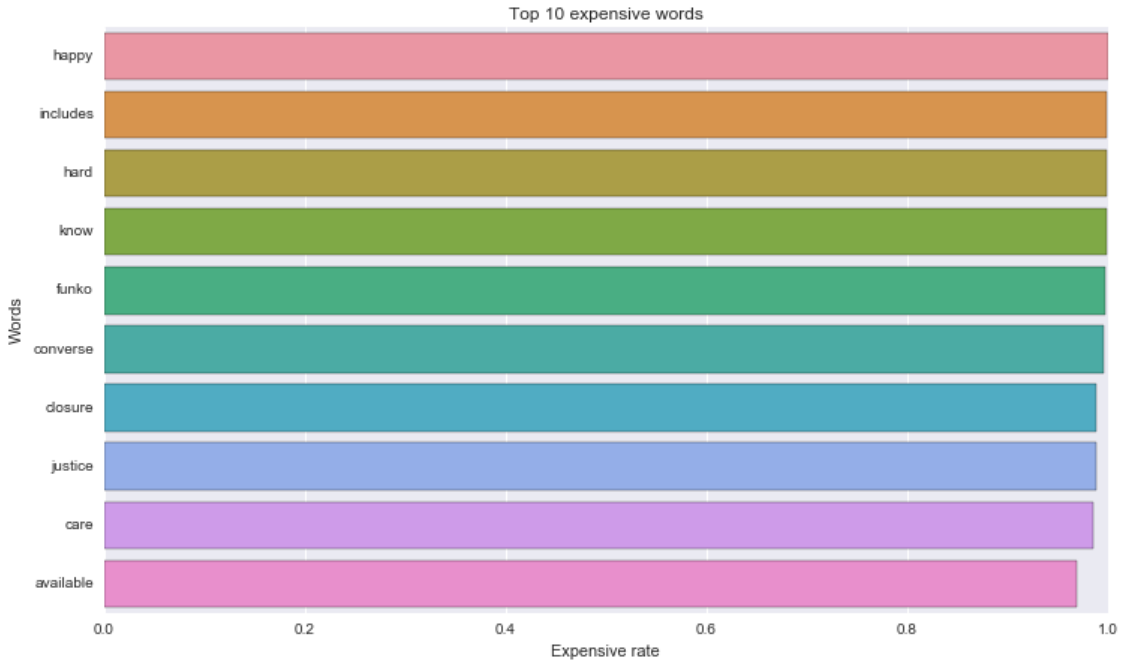




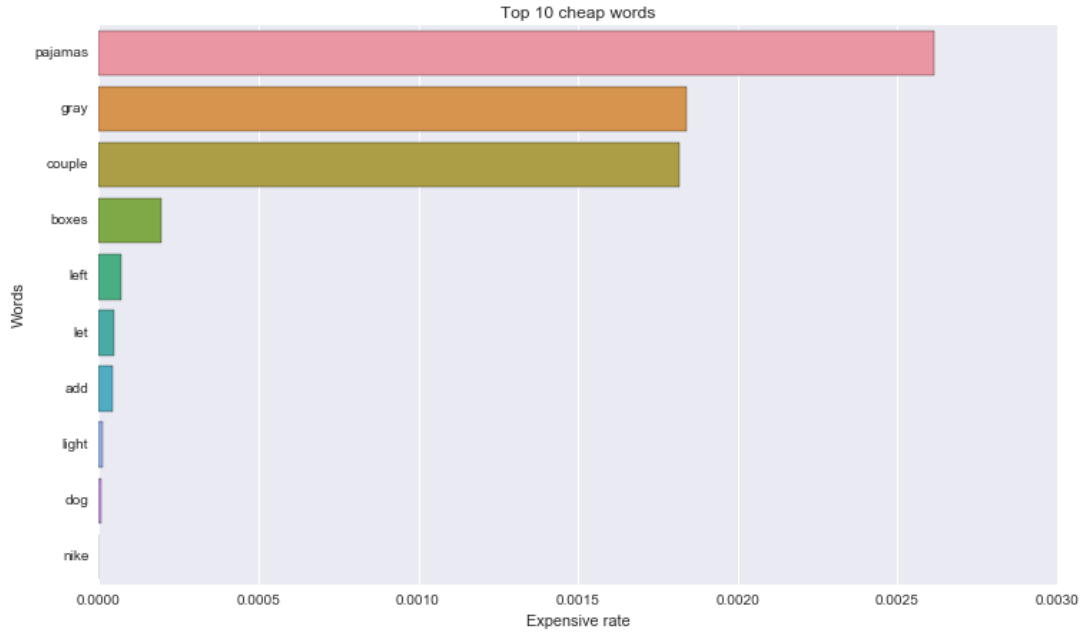


6.Item description rate:

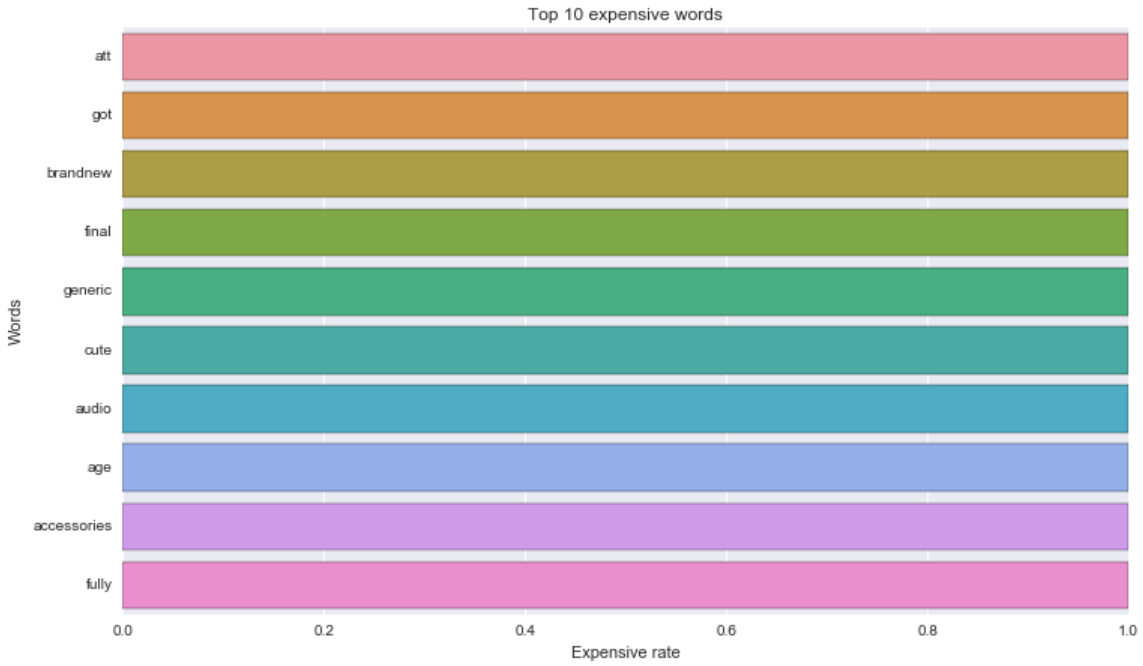
Expensive Words for “Woman” Category:



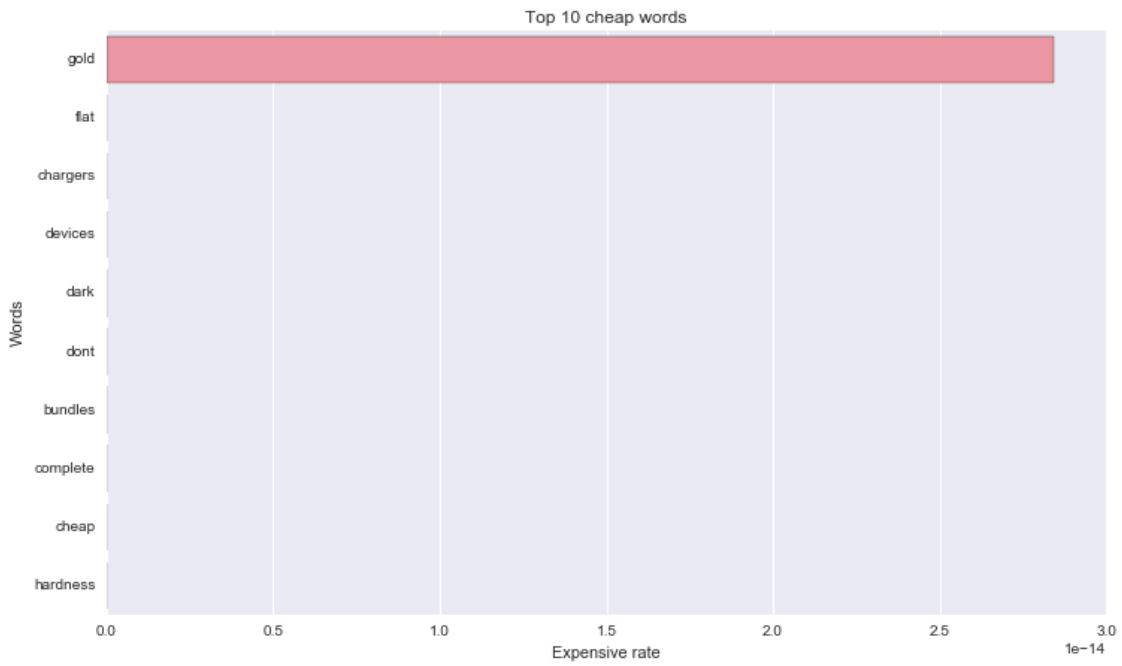
Cheap Words for “Woman” Category:



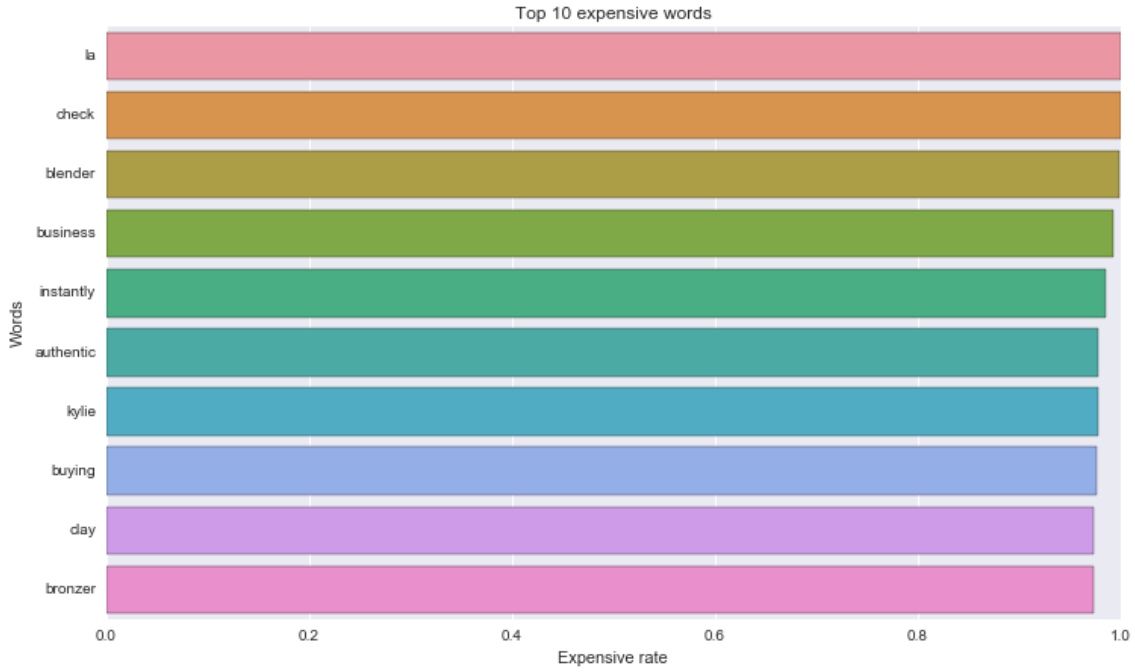
Expensive Words for “Electronic” Category:



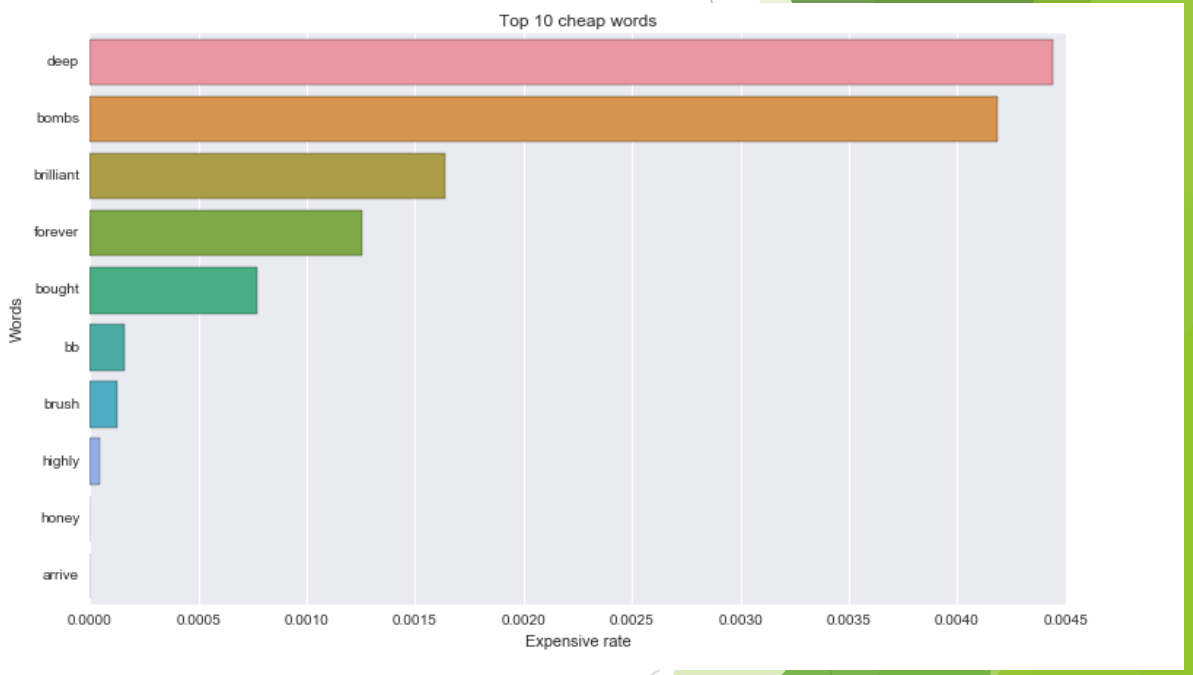
Cheap Words for “Electronic” Category:



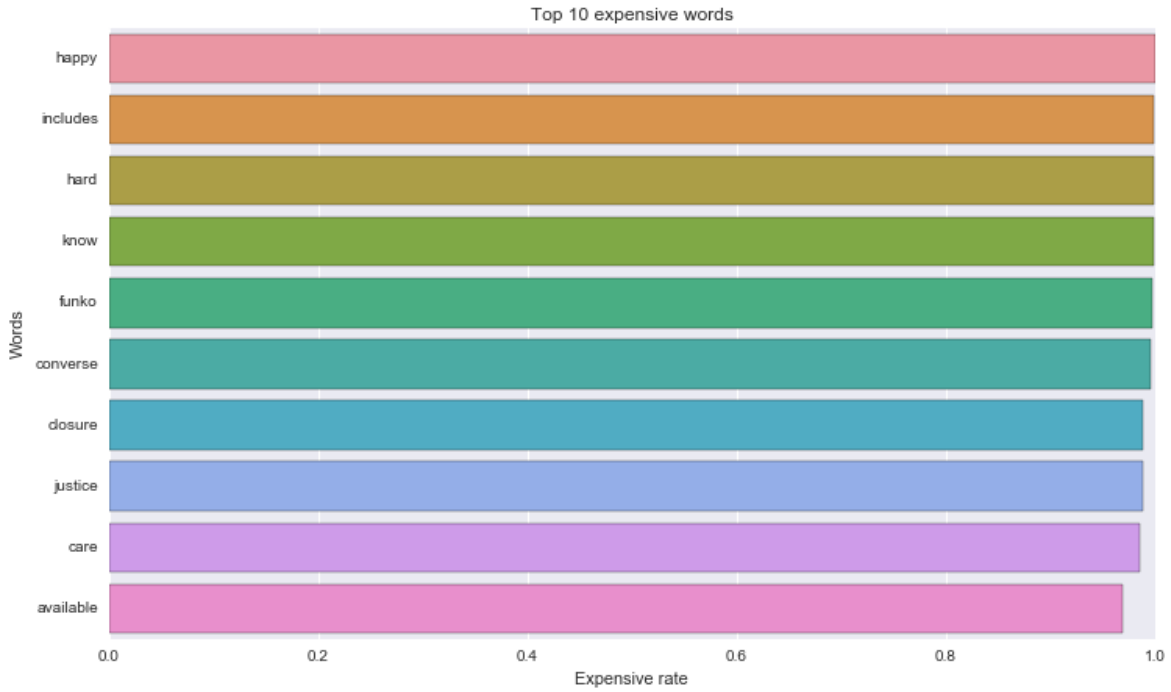
Expensive Words for “Beauty” Category:



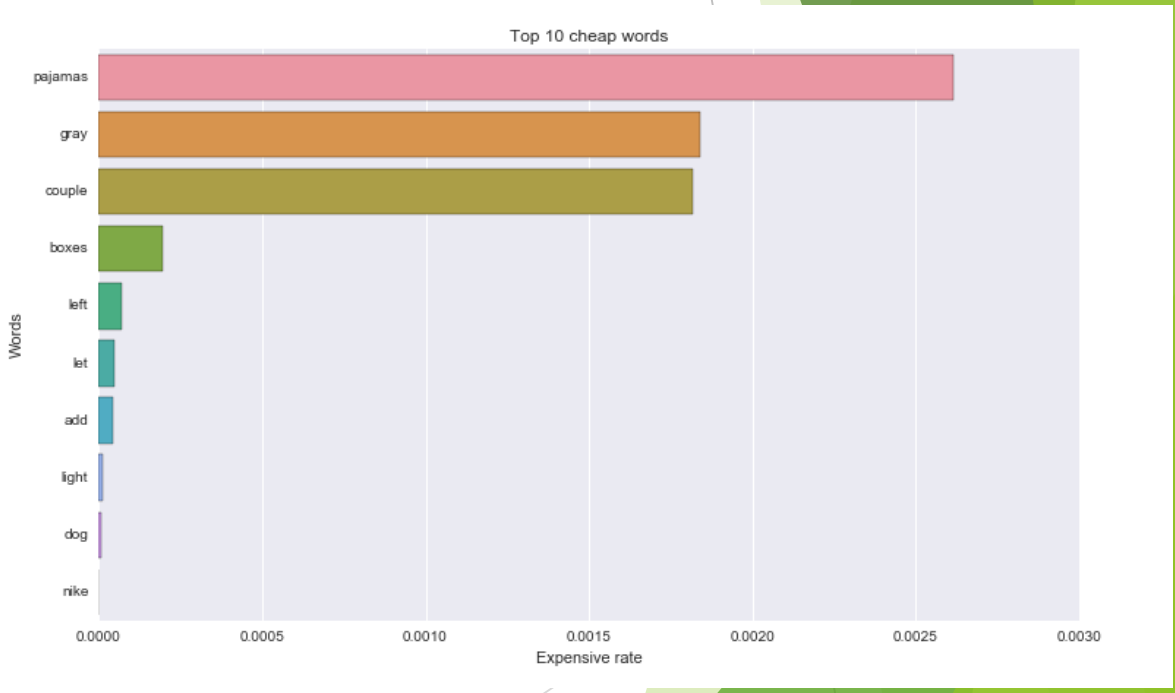
Cheap Words for “Beauty” Category:



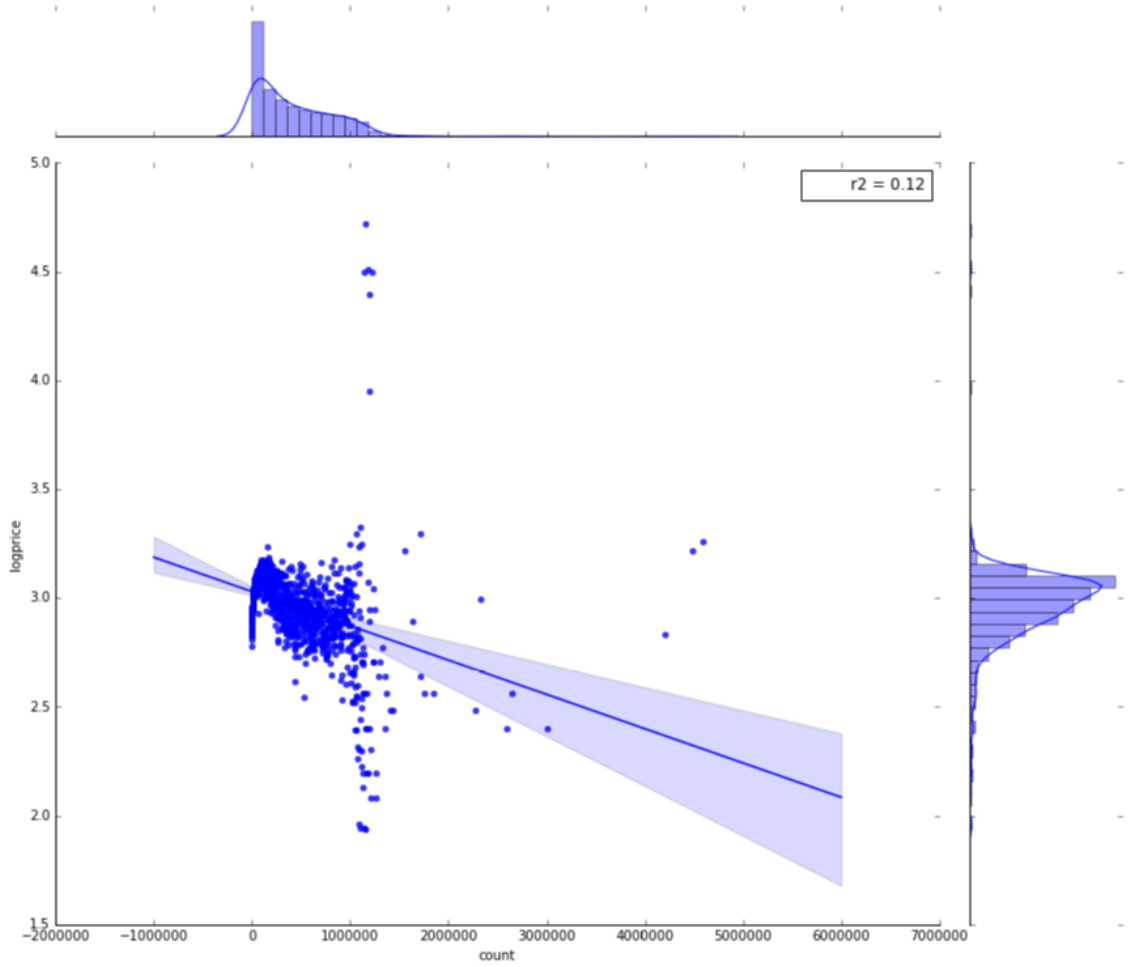
Expensive Words for “Kid Category:



Cheap Words for “Kid Category:



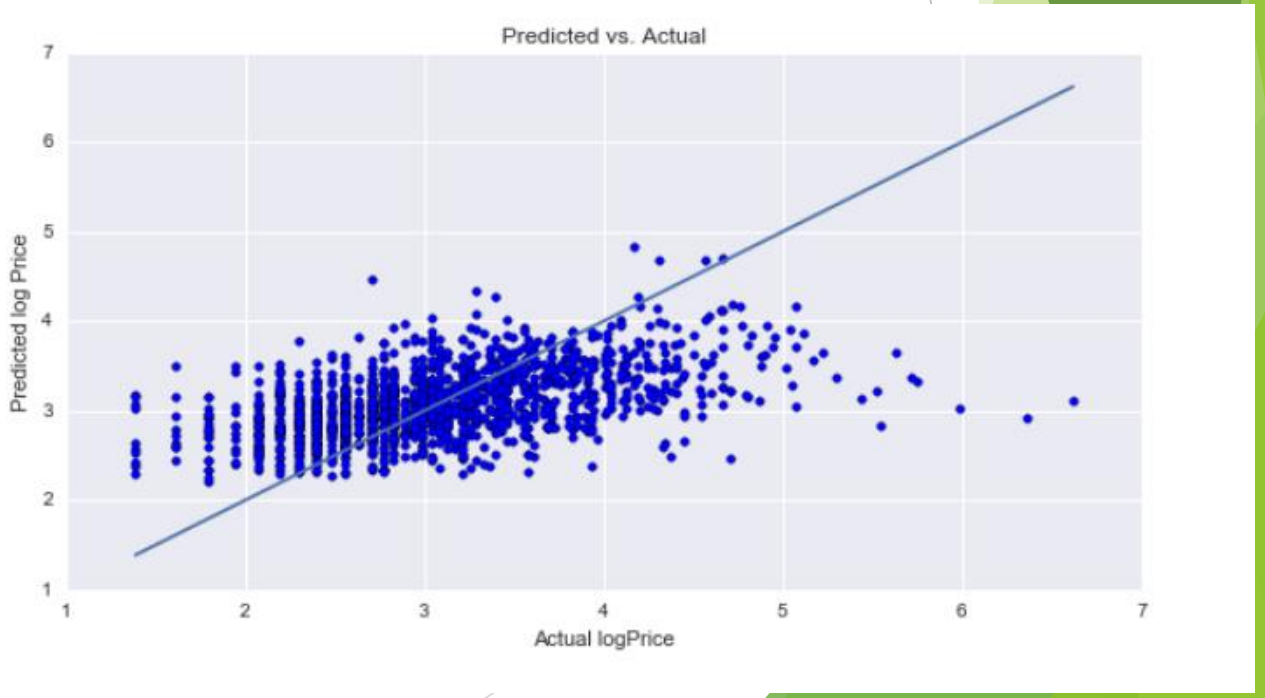
7. Quadratic Term of Word Count vs Log Price:



IV.Machine Learning:

Random forest :

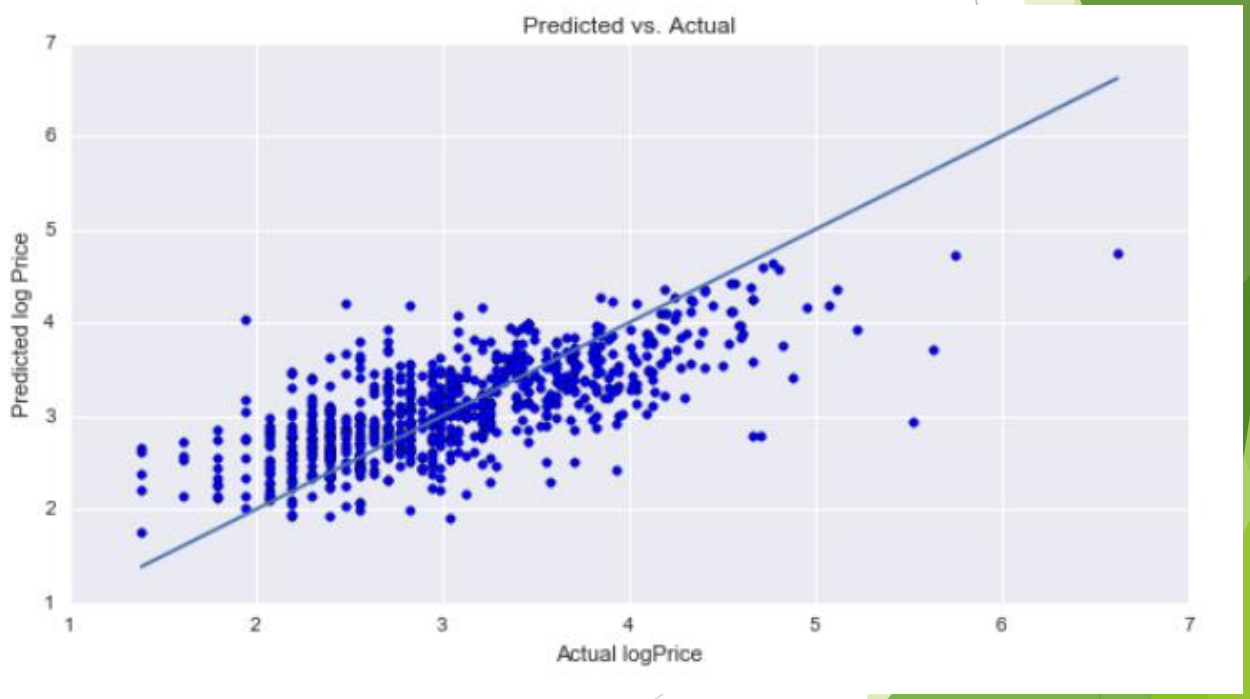
- R^2 for training set: 0.47
- R^2 score for testing set: 0.26
- Cross validation RMSE: 0.63



- Important features:

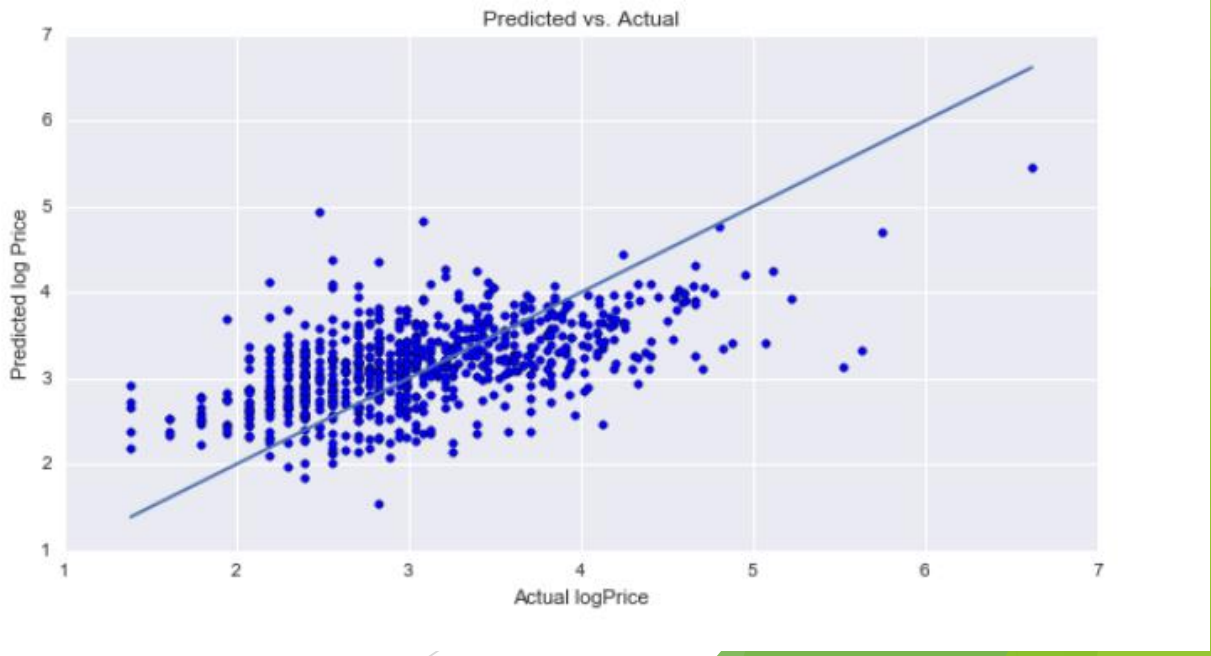
Ridge regression:

- R^2 score for training set :0.70
- R^2 score for testing set :0.49
- Cross validation RMSE : 0.56



SVR:

- R^2 score for training set :0.42
- R^2 score for testing set :0.26
- Cross validation RMSE : 0.65



V. Conclusion:

This project has opened up my mind into the knowledge of NLP and it showed me how much pre-processing steps are involved to analyze text data. I learned the most common steps for text pre-processing and the choice of algorithms and how important computation is when you're dealing with large datasets. Through this project, I selected ridge regression was the best method in order to predict price. Mecarri can use my analysis in order to offer a suggestion price for the sellers by item detail and category . However, the weakness of this analysis could not predict price for all items due to memory error, but will try to learn other methods to optimize my analysis in future.