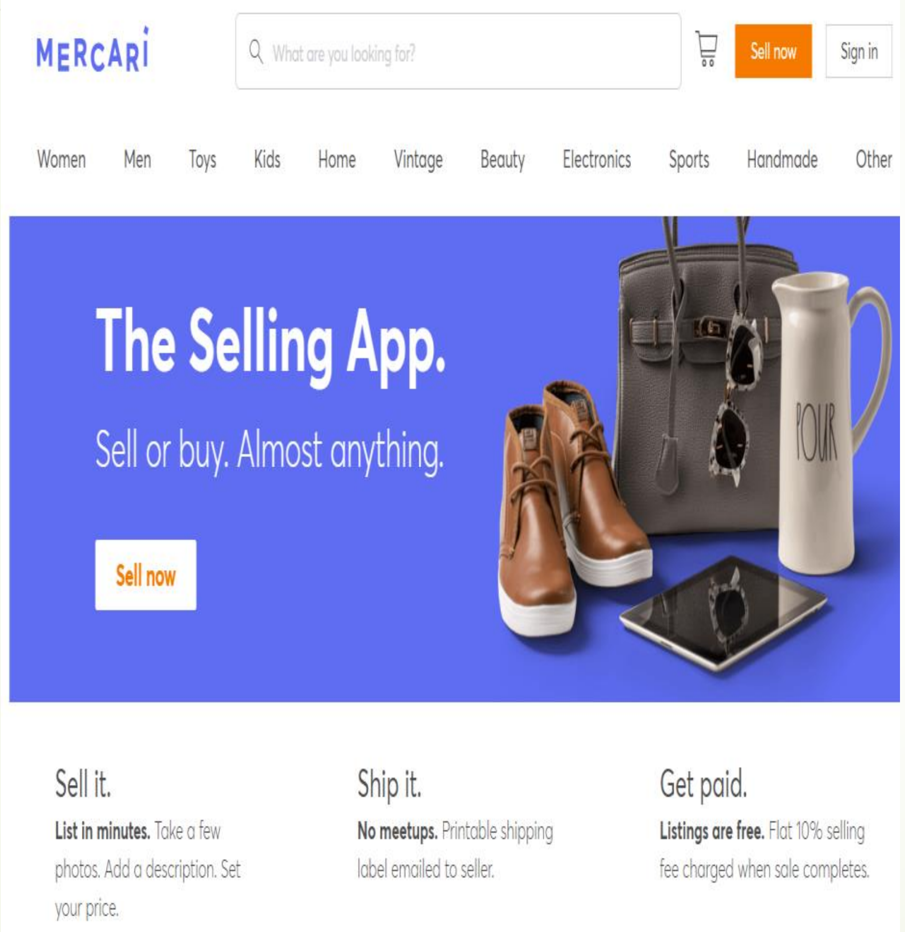
Mercari Price Prediction

Data Science project by Laxmi Vanam

Overview

- The Business Problem
- Dataset Features
- Evaluation metric
- Exploratory Data Analysis
- Text Processing
- Vectorization
- Combining features and modeling
- Future Enhancement

The Business problem



The Selling App.

Sell or buy. Almost anything.

Sell it.

List in minutes. Take a few photos. Add a description. Set your price.

Ship it.

No meetups. Printable shipping label emailed to seller.

Get paid.

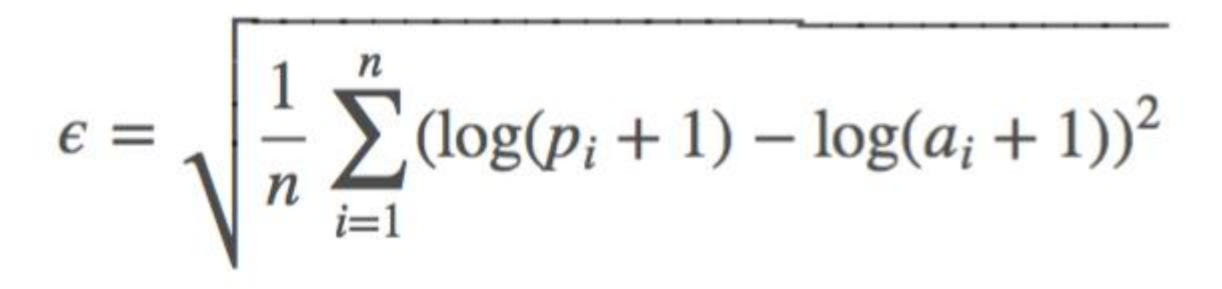
Listings are free. Flat 10% selling fee charged when sale completes.

- Mercari is a marketplace, where users can upload products to see it online.
- The challenge is to create a model that would help sellers price the product.

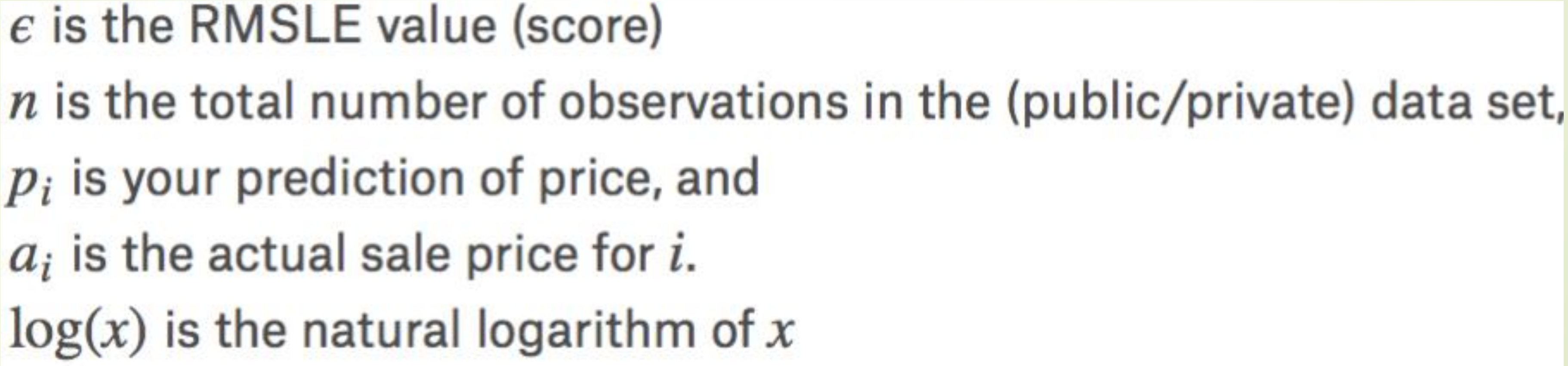
Dataset features

- **ID**: the id of the listing
- **Name:** the title of the listing
- **Item Condition:** the condition of the items provided by the seller
- **Category Name:** category of the listing
- **Brand Name:** brand of the listing
- **Shipping:** whether or not shipping cost was provided
- **Item Description:** the full description of the item
- **Price:** the price that the item was sold for. This is the target variable that you will predict. The unit is USD.

Evaluation metric



$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_i + 1))^2}$$



ϵ is the RMSLE value (score)

n is the total number of observations in the (public/private) data set,

p_i is your prediction of price, and

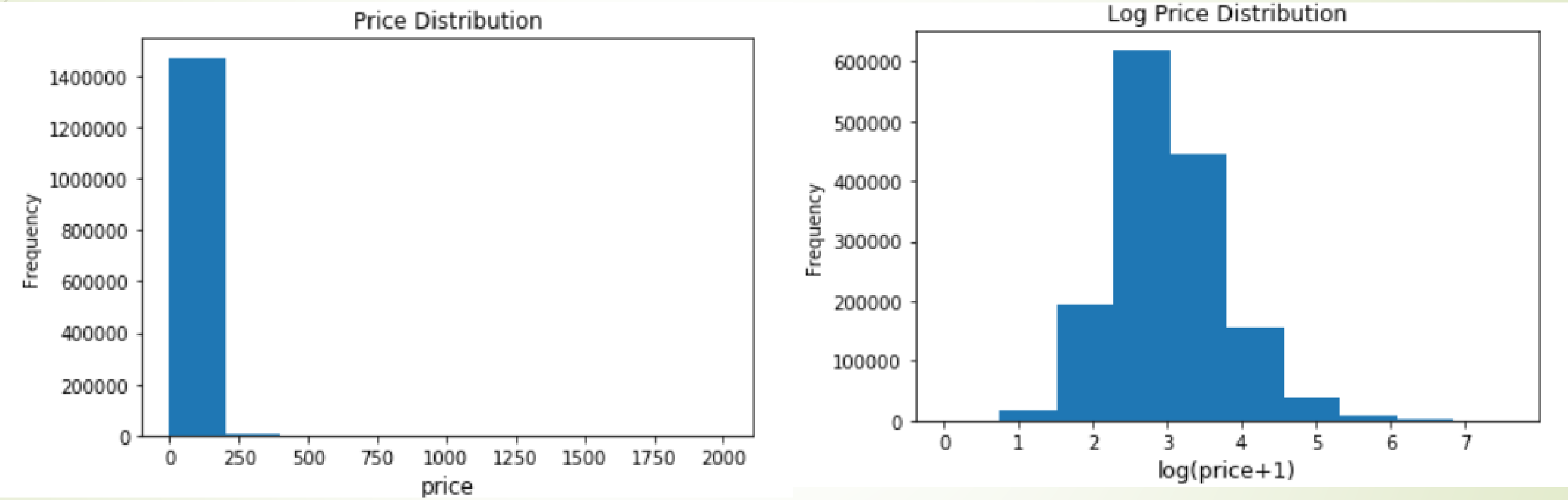
a_i is the actual sale price for i.

$\log(x)$ is the natural logarithm of x

Exploratory Data Analysis

- Observe Training Statistics
- Simple Data Inspection
- Handling missing values

Exploratory Data Analysis contd..

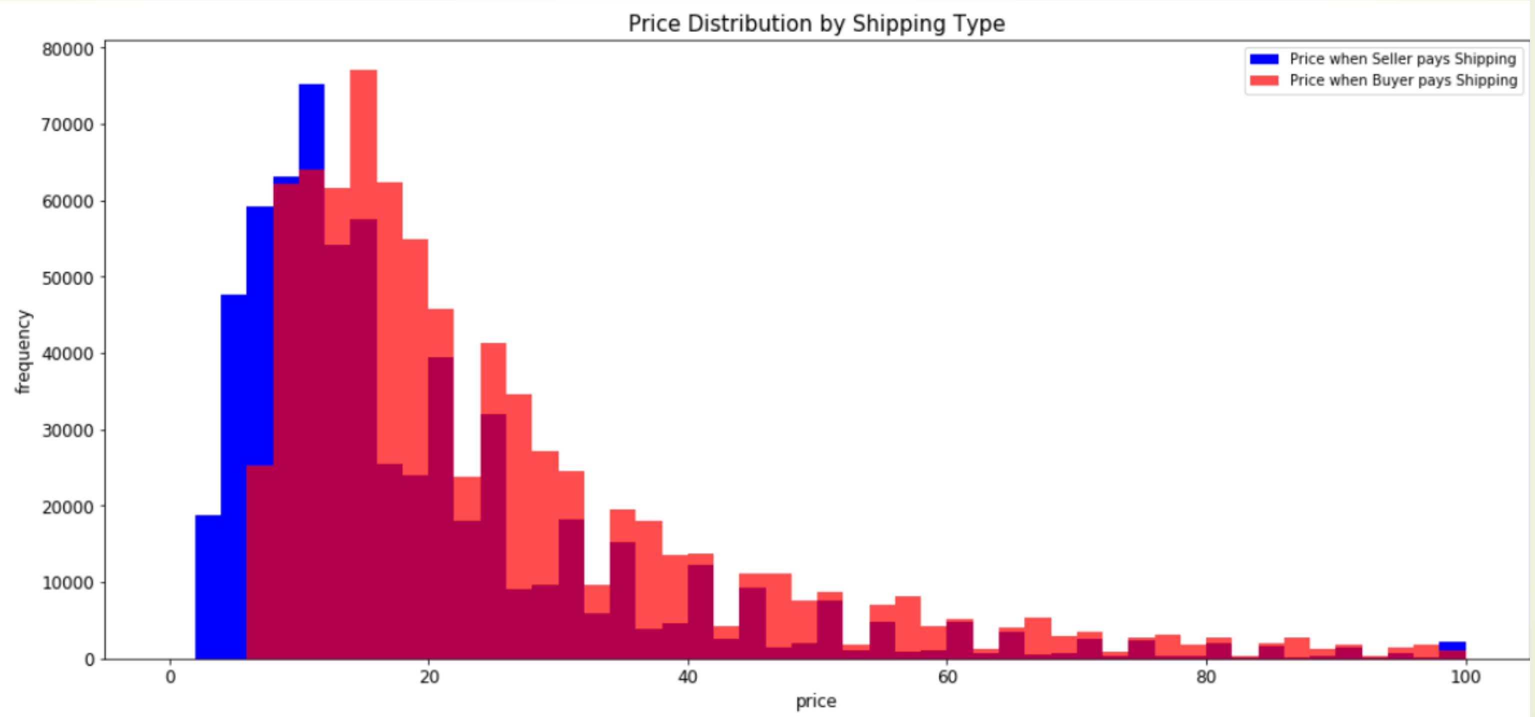


Price Distribution

Log Price Distribution

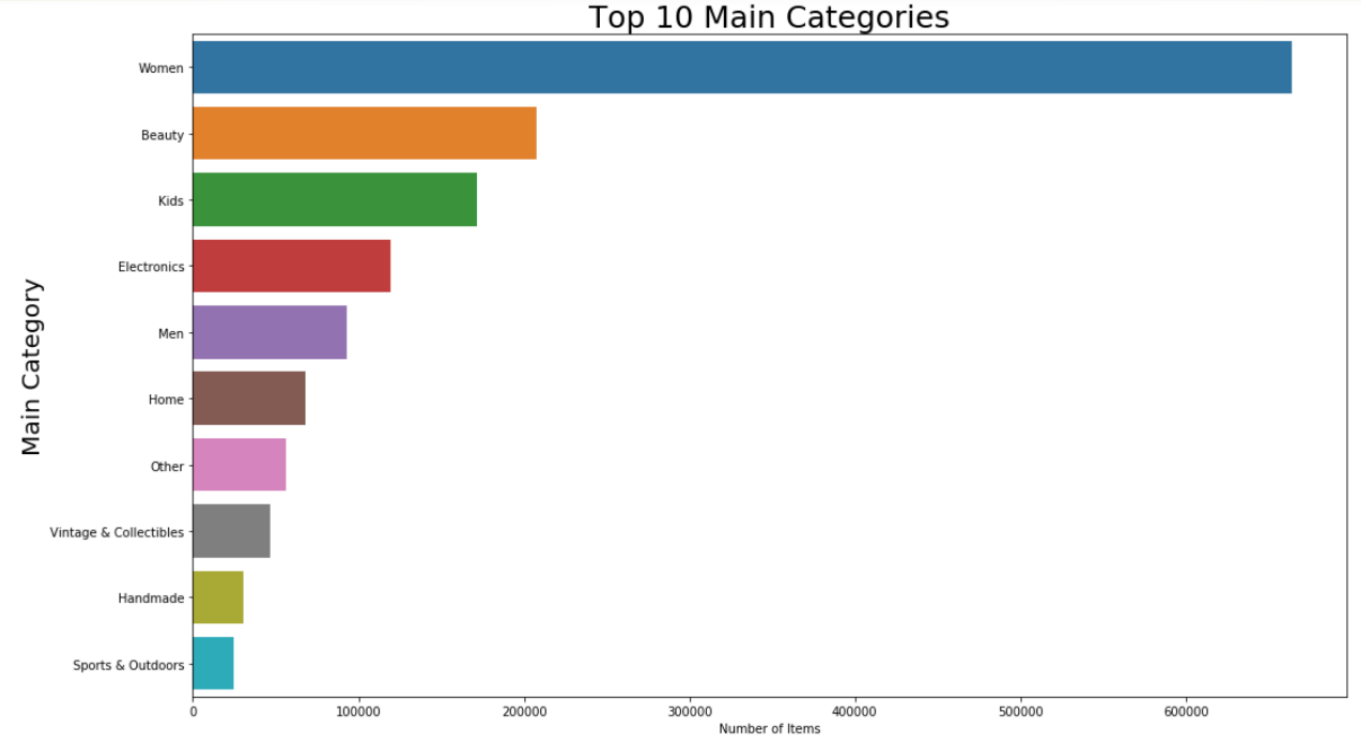
Distribution of the price and its log value

Exploratory Data Analysis contd..



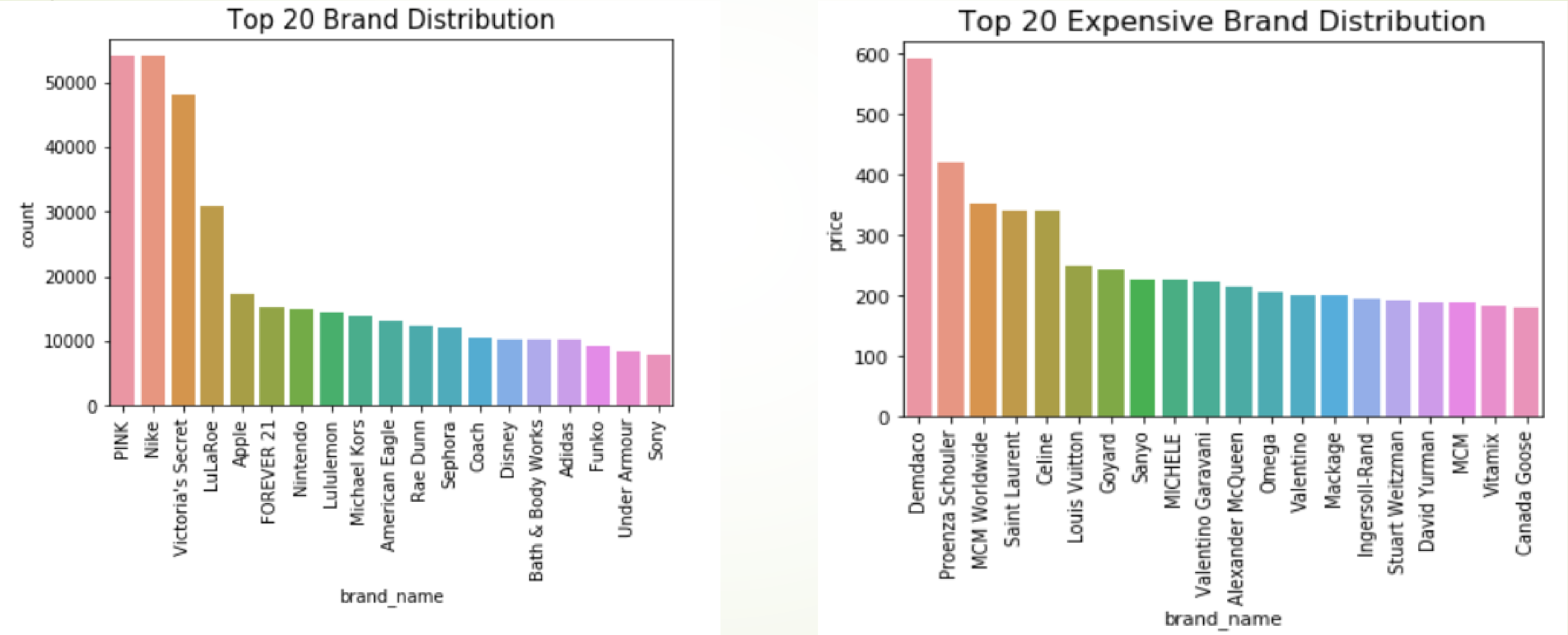
Distribution of the by shipping

Exploratory Data Analysis contd..



Distribution of Category

Exploratory Data Analysis contd..

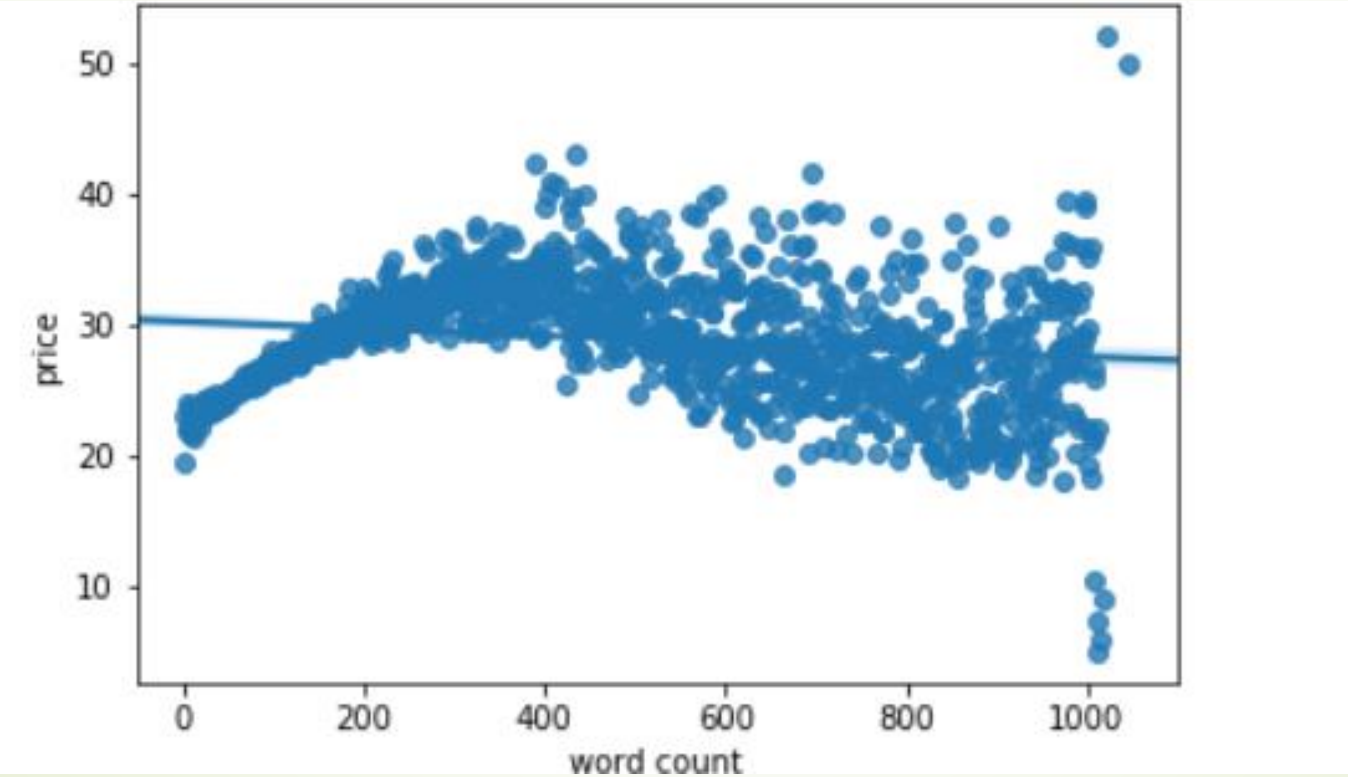


Top 20 Brand Distribution

Top 20 Expensive Brand Distribution

Distribution of the Brand

Exploratory Data Analysis contd..



Item description length by price

Text Processing

- Normalization
  - Removing Punctuations/ Stop Words/ lowercasing the words/ Stemming or lemmatizing the words etc.
- Tokenization/ Bag of words modeling
  - Using Ngrams to preserve local ordering of words to improve model performance.
- Vectorization/ Scoring words
  - Reducing text to a vector using CountVectorizer/TF-IDF/LabelBinarizer

Vectorization

- CountVectorizer:

  Returns an encoded vector with integer count for each word

- TF-IDF:

  This is to capture rarity of the word. This is to find frequent terms from the document that isn't so frequent within the whole document corpus.

- LabelBinarizer:

  Get's all the word and assigns it to its own column. 0 means it's there and 1 means not (example with brand names)

Combining features and modeling

- Handling sparse matrices
- Train test split
- Modeling using Keras regression
- Prediction

Possible Enhancements:

- Do more feature engineering to come up with more features.
- Try more modeling techniques and tune them for better metric.
- Use decomposition techniques in order to reduce the dimensions.