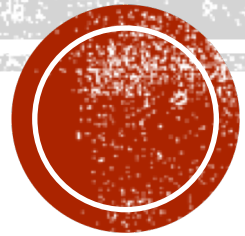
MERCARI PRICE SUGGESTION PROJECT

Student: Nathan Zhang

Advisor: Misael Manjarres



CONTEXT

- Mercari: Japan’s biggest community-powered shopping app
- Mercari offers price suggestions based on descriptions from users

Sweater A:

“Vince Long-Sleeve Turtleneck Pullover Sweater, Black, Women's, size L, great condition.”

Sweater B:

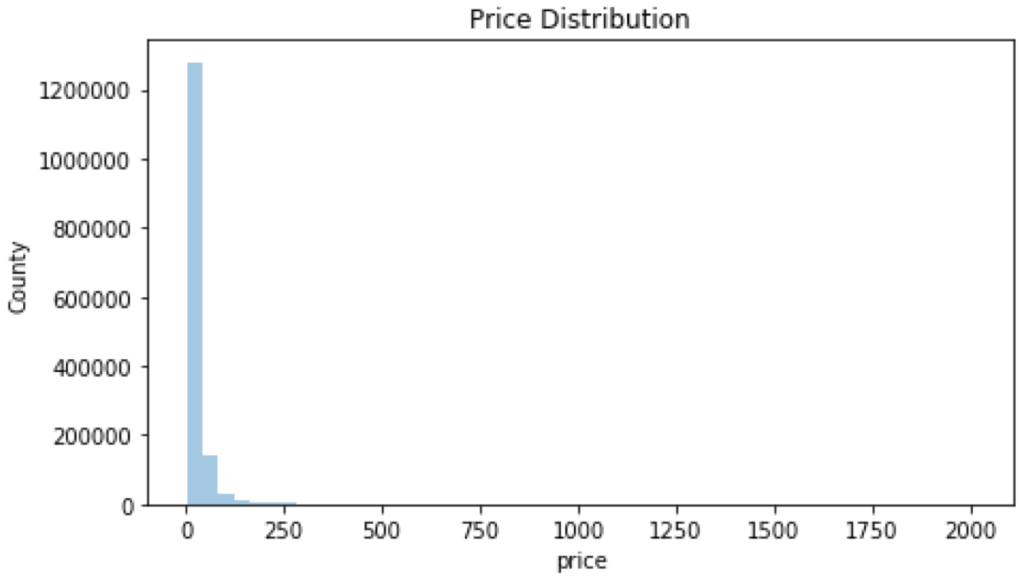
“St. John's Bay Long-Sleeve Turtleneck Pullover Sweater, size L, great condition”

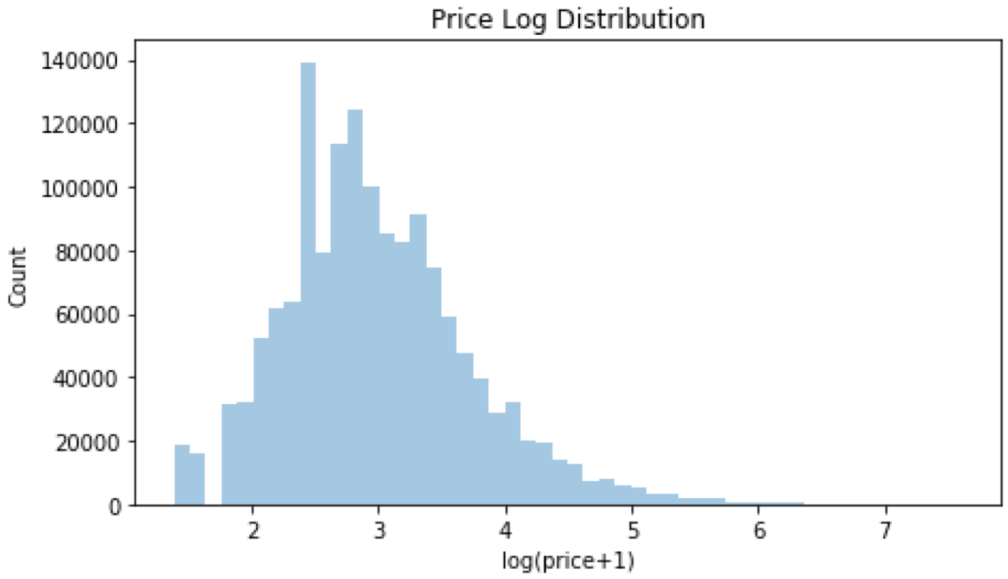
DATA

- Features
  - Name: title of the listing
  - Price: numerical ($3-$2000)
  - Item condition
  - Category: general, sub, sub 2
  - Brand: ~ 600K missing
  - Shipping cost: whether covered by seller
  - Item Description: user-written description of the item
- Size
  - Train: ~1.5M
  - Test: ~700k

EDA & FEATURE ENGINEERING

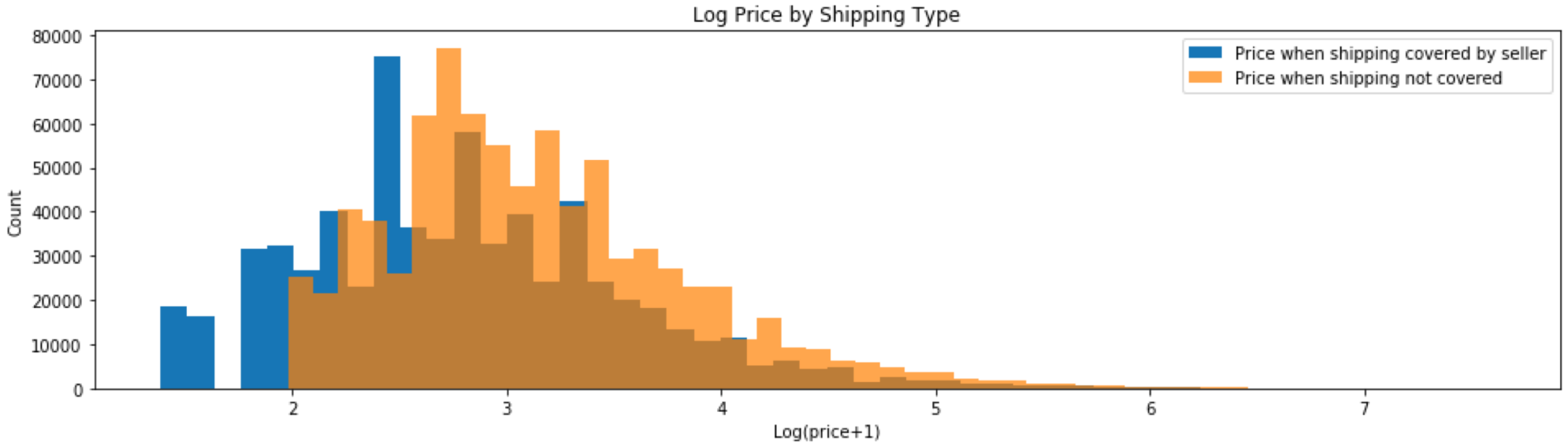
- Price -> Log Price





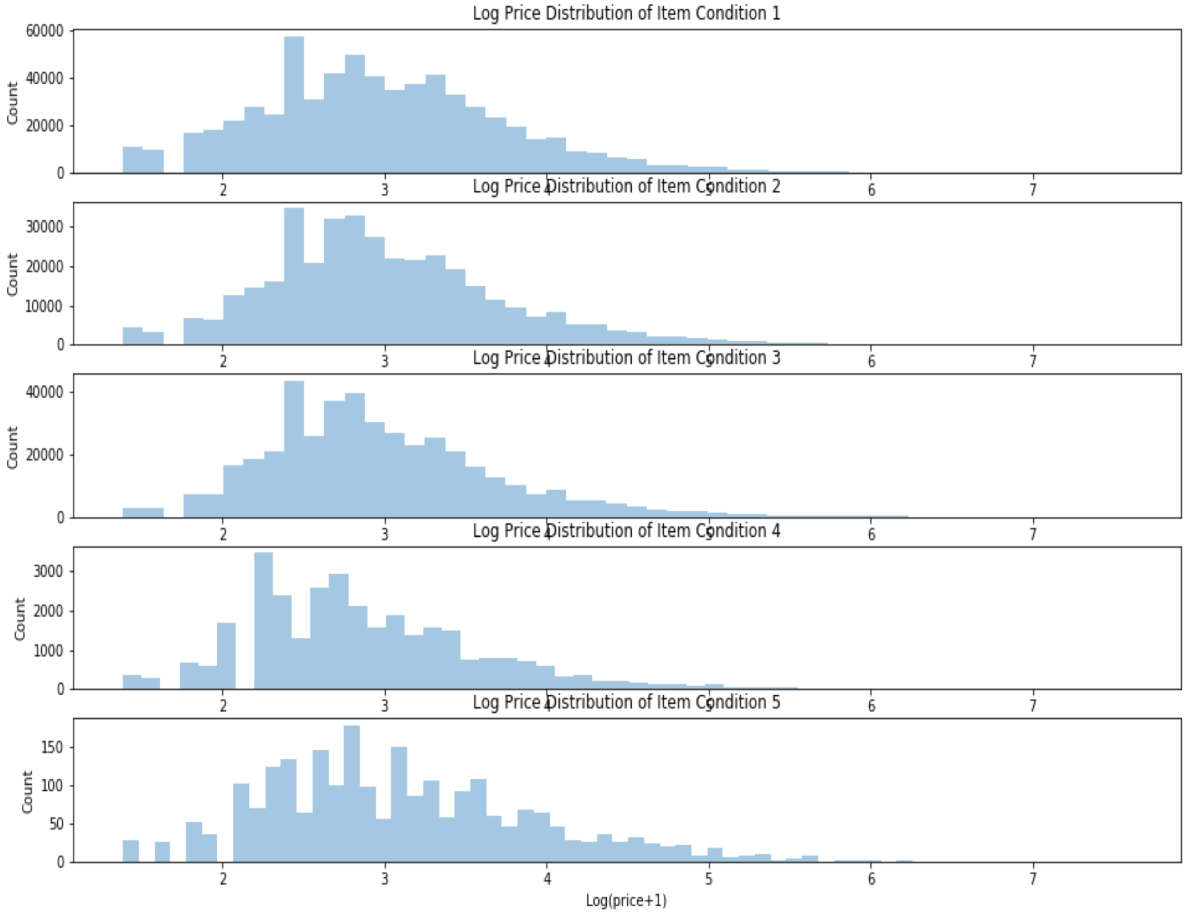
EDA & FEATURE ENGINEERING

- Shipping Cost
  - 45% sellers pay for shipping
  - Shifted similar-shape distribution (due to addition of shipping cost)



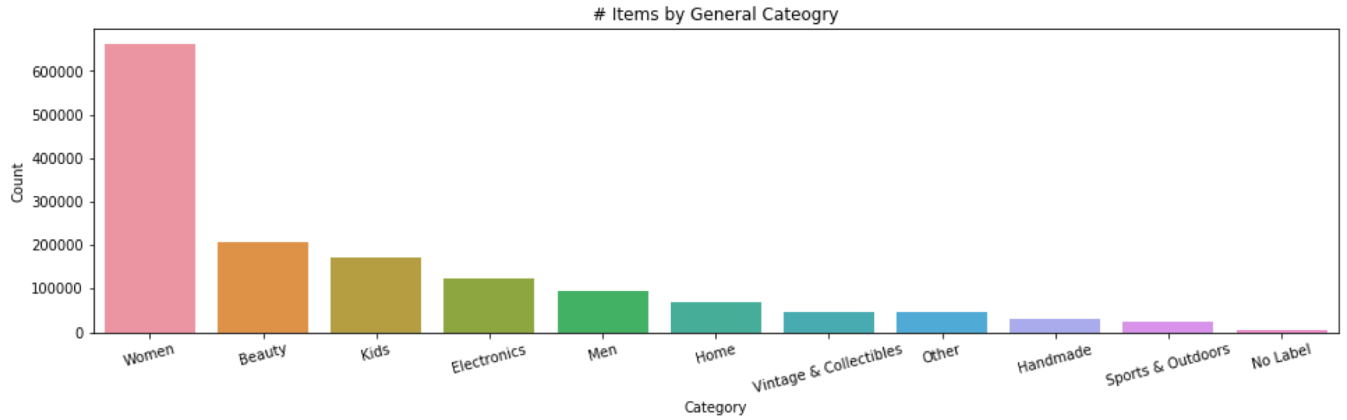
EDA & FEATURE ENGINEERING

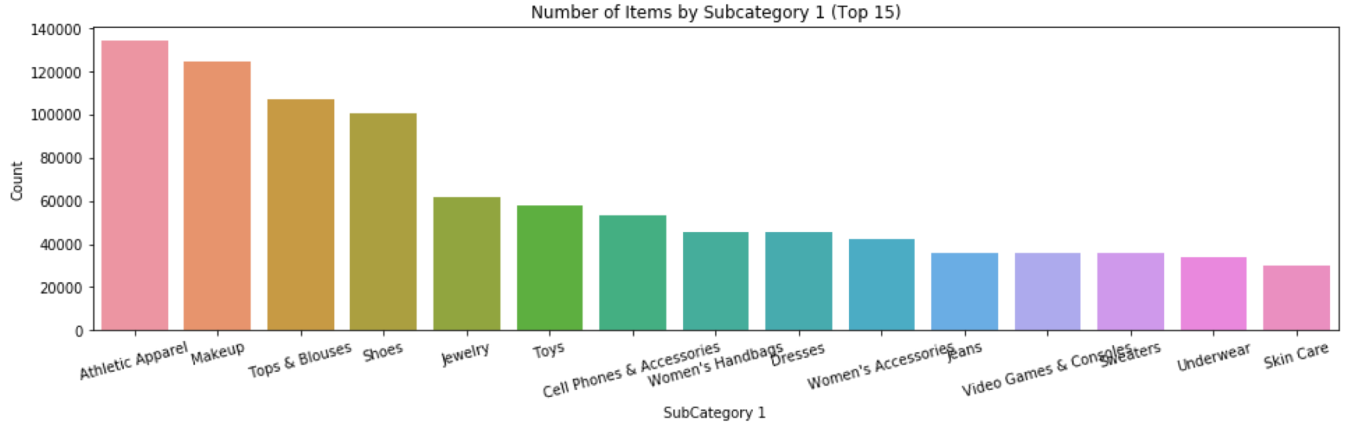
- Item Condition
  - Indicated as 1, 2, … , 5
  - Large variance
  - Similar shape



EDA & FEATURE ENGINEERING

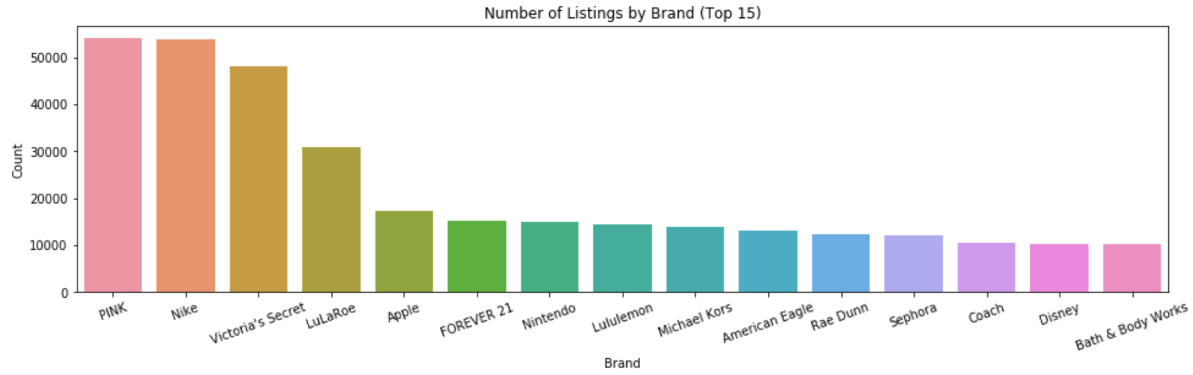
- Item Category
  - ~1300 categories
  - Largest category: Women

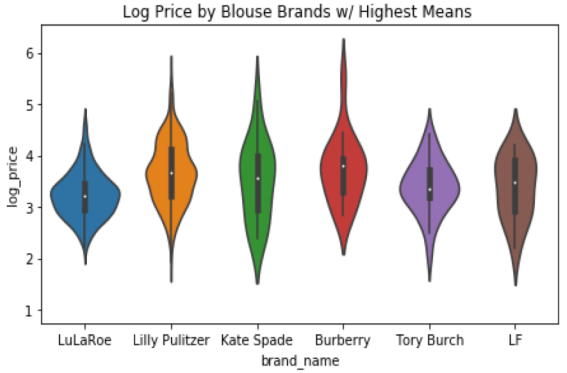




EDA & FEATURE ENGINEERING

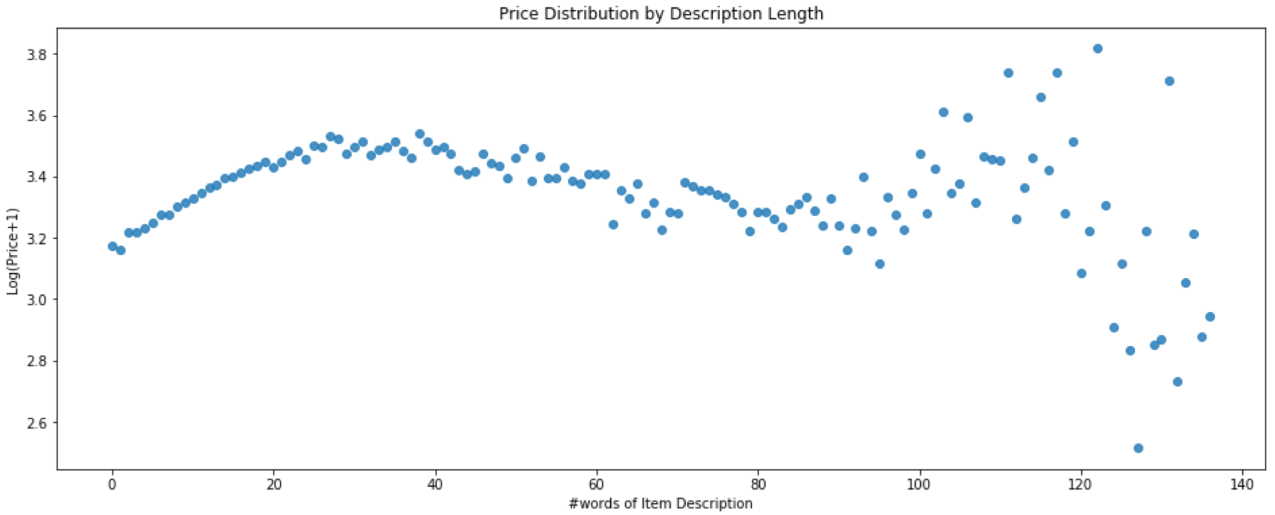
- Brands
  - ~4800 brands
  - Largest category: Women
  - Each category has different brands
  - Products of same brand in different categories (e.g. Disney, VS, Nike)





EDA & FEATURE ENGINEERING

- Item name & description
  - User written
  - Lengthy description higher price
  - May contain brands that were missing
- Text process
  - Tokenize (punctuation, short, stop words)
  - Merge all text
  - TF-IDF
  - Label Encoding
  - Visualization (t-SNE)



MODELING

- Preprocessing
  - Convert categorical variables into dummies
  - Merge vectors with dummies
- Clustering (K-means)
  - Minibatch K-means
- Neural Network
  - # hidden layers, # nodes
- Ensemble method
  - Mean from KM and NN

CONCLUSION & MORE

- Not yet finished
  - long processing time
  - kernel kept dying
- TF-IDF improvement
  - Topic based (Latent Dirichlet Allocation)
- Other models
  - Forest Regression
  - XGBoost