

Applied Machine Learning: Final Project Deliverable 1
Data Analysis and Visualization

Mercari Price Prediction

Team 27

Arnav Saxena (as6456)

Nuanyu Shou (ns3492)

Oliver (Hongou) Liu (hl3518)

Xinfu Su (xs2444)

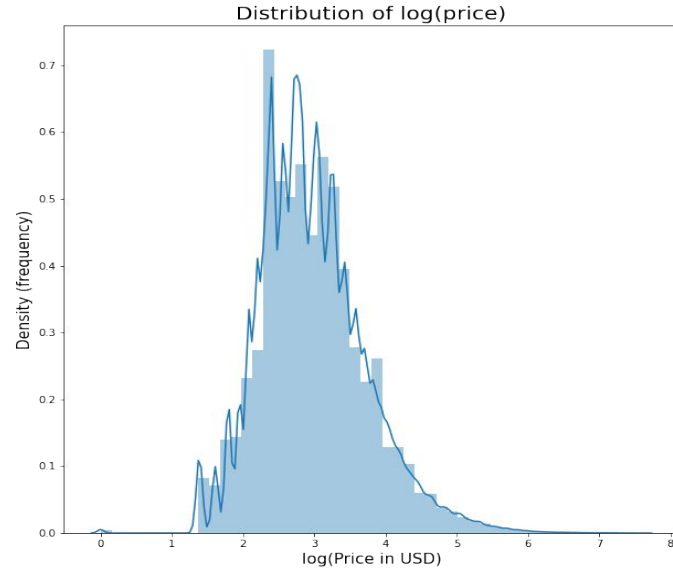
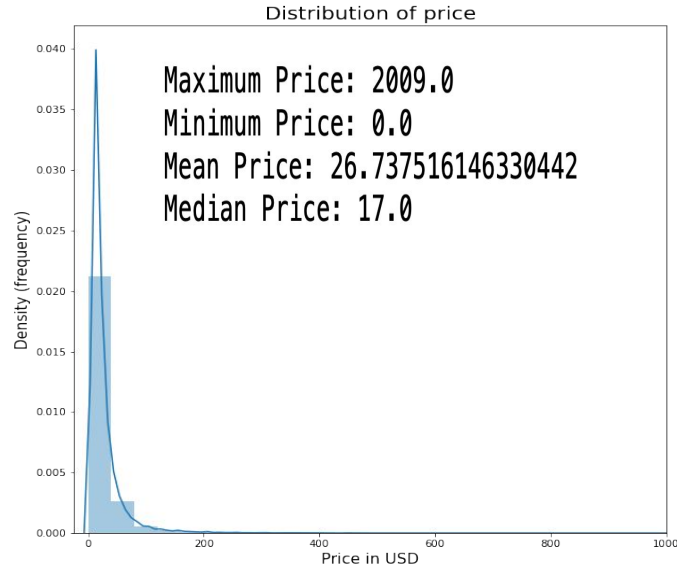
Zihao Liu (zl2986)

- **7 key features** in total

- What we have done:
 - Exploration 1: **distribution** of one single feature
 - Exploration 2: relationship **between each feature and target**
 - Exploration 3: **interaction between two variables** and their impact on target

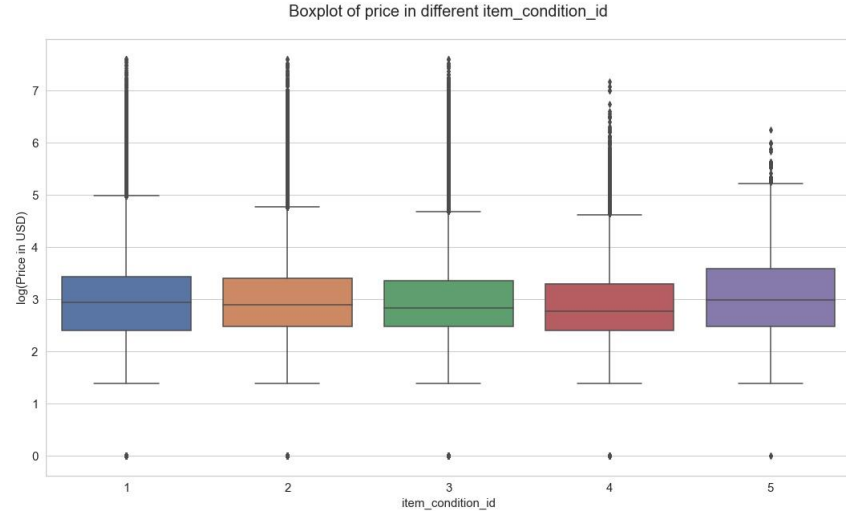
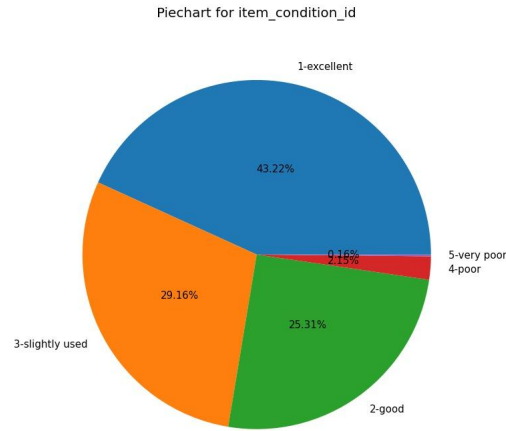
- Motivation of EDA:
 - To determine how we should **pre-process each feature** if needed. (e.g. log transformation on price)
 - To determine what **encoding method** we should use.
 - To determine if any **interactive feature engineering** (e.g. interaction term) should be applied.
 - To determine what models to use in our future work

Analyzing Target variable - Price



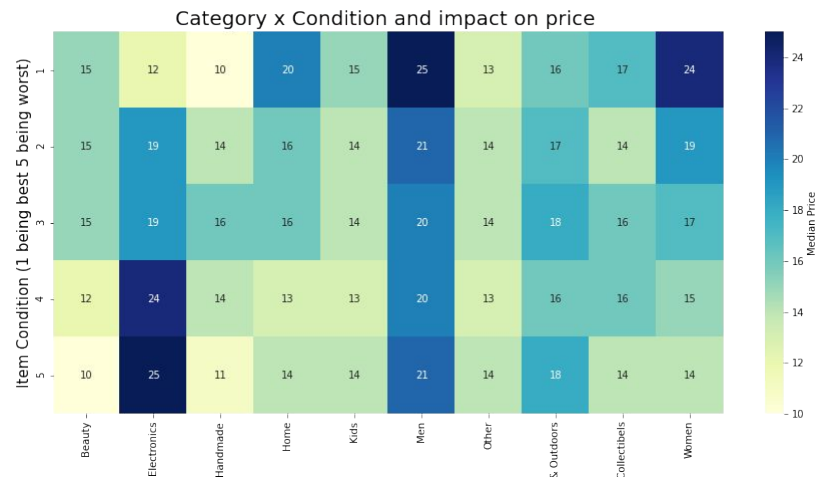
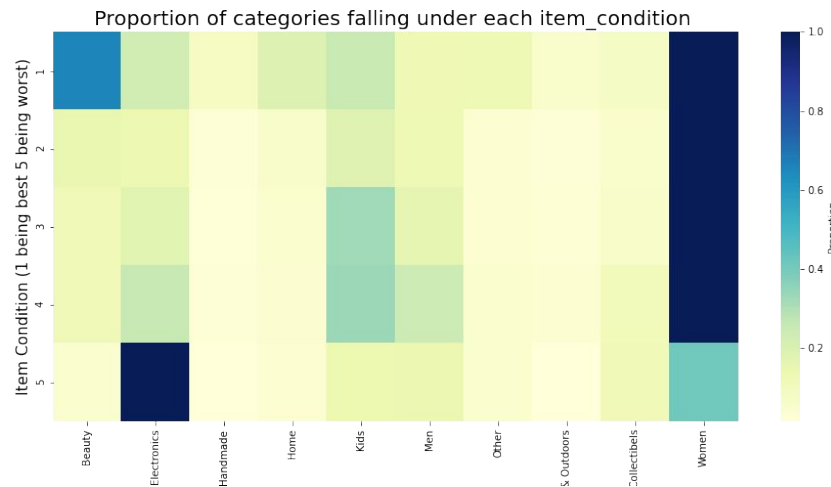
- The histogram on the left shows that **price variable is highly positively skewed**. The maximum price goes up to \$2009 where as the mean and median hovers around ~\$27 and \$17 respectively.
- Hence we decided to **apply log transformation to price + 1** (+1 since some items had zero price which was resulting in an error during log transformation)

Analyzing item_condition_id



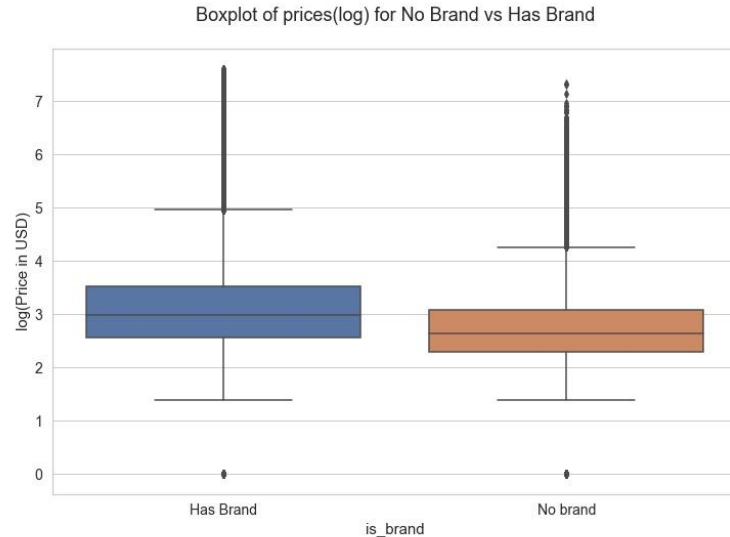
- The pie chart on the LHS suggests that the distribution of different “item_condition_id”s is skewed towards good condition. As can be expected, **very few items (<2.5%) are being marked under “poor” condition (id 4 & 5)**
- An interesting observation we note while studying the relationship between item condition and price was that the **median price of items gradually decrease as we move from condition 1(good) to 4 (poor)**. However, quite unexpectedly, we find that the median price of items lying under the very poor condition (id 5) is highest. Hence, ordinal encoding shouldn't be used.

Analyzing item_condition_id (cntd...)



- Using the heatmap on LHS we note that **condition 5 items majorly comprise of electronics**. The RHS further shows that for some reasons **electronics with poor conditions (id 4 & 5) are the most expensive** items of the lot (as they might include heavily used but longer shelf life items such as laptops etc). This **explains why the median price of items with the most used conditions was the highest** in the box plot on last slide
- Since ordinal encoding is out of question (as explained in previous slide) - we can **experiment with categorical and target encoding for “item_condition_id”**

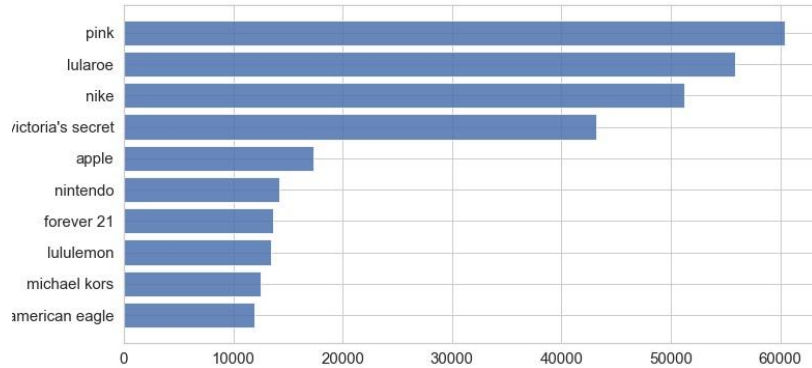
Analyzing brand_name



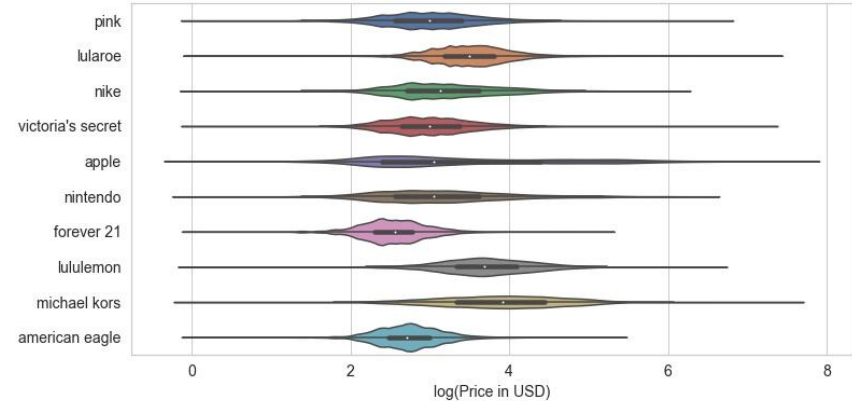
- Initially we **had ~42% observations with missing “brand_name”**. But we observed that most sellers have put the brand name in the “item_name” column. Thus we **extracted the brand names from the item name** column and **reduced missing “brand_names” to ~28%**
- We further found that **items that have an associated brand name have higher median price** as compared with items with no brand names. Hence we decided to treat the missing values as a new category under “brand names”.

Analyzing brand_name (brand_name x price)

Ten most popular brand names and counts




Violinplot of price(log) for 10 popular brands



- There are **4720 different brand names in the dataset**. We pick the 10 most popular brands and draw the violin plot between price and brands. As expected, we find that **brands have a key relationship with the price of the product**
- Keeping in mind the number of categories and the impact of brands on price levels, we **perform target encoding on the “brand_name” column**

Analyzing category_name

category_name
Beauty/Makeup/Lips
Electronics/Cell Phones & Accessories/Cases, C...
Women/Tops & Blouses/Blouse
Home/Kitchen & Dining/Coffee & Tea Accessories
Beauty/Tools & Accessories/Hair Styling Tools

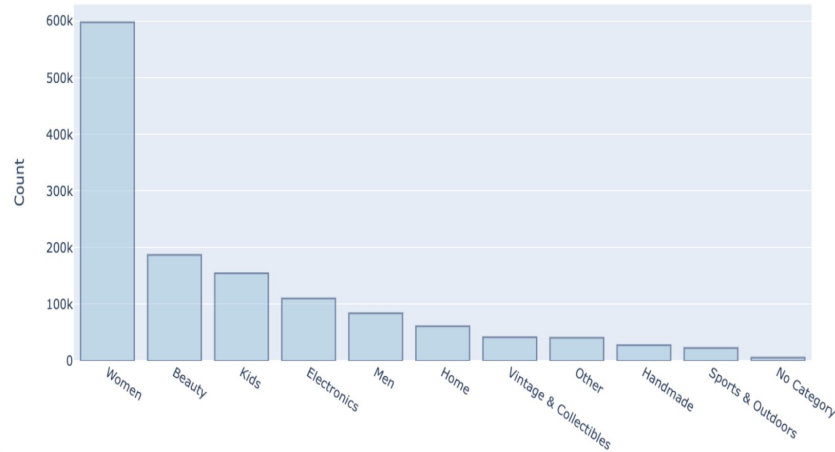


main_cat	subcat_1	subcat_2
Beauty	Makeup	Lips
Electronics	Cell Phones & Accessories	Cases, Covers & Skins
Women	Tops & Blouses	Blouse
Home	Kitchen & Dining	Coffee & Tea Accessories
Beauty	Tools & Accessories	Hair Styling Tools

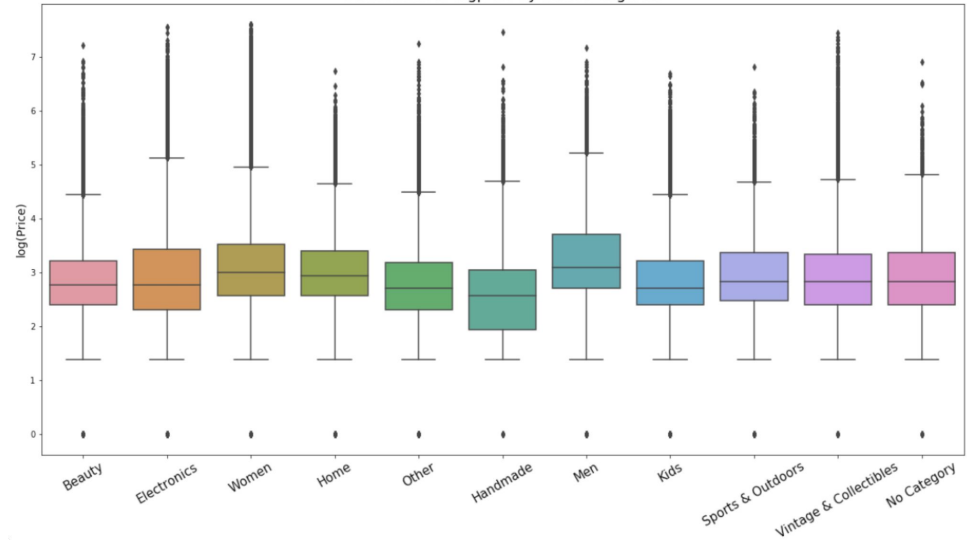
- The “category_name” contains values as:
main category/sub_category_2/sub_category_3.
Hence, we split the column into three different columns
as shown in the figure in LHS

Analyzing category (main_category x price)

Number of Items by Main Category



Distribution of logprice by main categories



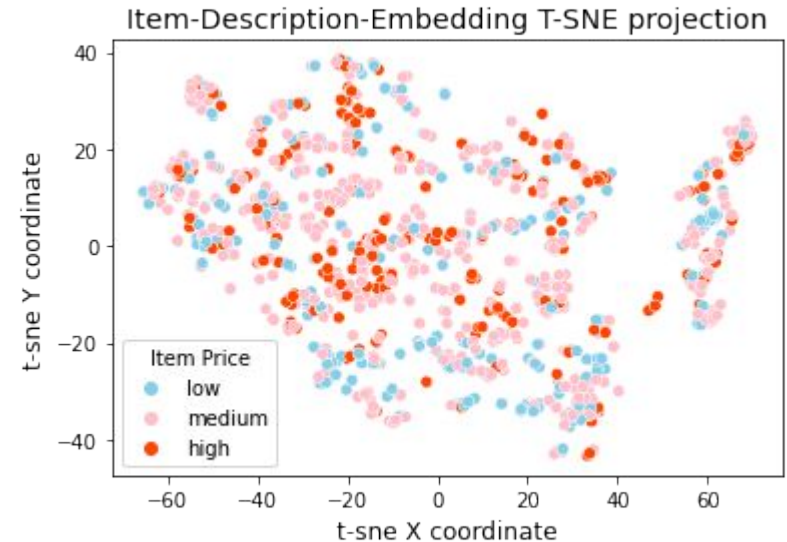
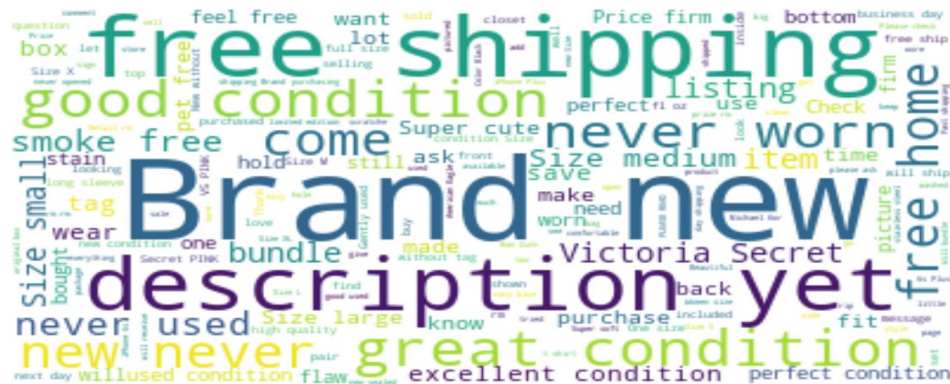
- The bar chart shows **high imbalance in the distribution of main category**. 'Women' products account for the largest number of categories followed by beauty at a distant second.
- The box plot on RHS, quite unexpectedly shows that price of the items isn't really determined by the main category. This indicates that the **category column by itself isn't really impacting the price of products** (for instance maybe electronics includes expensive laptops as well as inexpensive christmas lights)
- Thus in order to efficiently capture the interaction between categories and sub_categories, we will encode the **main_category categorically** and **perform target encoding on the sub categories**

Analyzing shipping



- As suggested by the boxplot in the middle - items where **shipping is paid by buyer** tend to have **higher median prices** as compared to those where the seller pays for shipping
- The heat map on the RHS shows that **buyer are asked to pay for shipping of items which are in good conditions** while sellers are more indexed to paying for shipping for items that aren't in that good of a shape (probably in a bid to make up for the poor condition of the item). As already discussed in slide 4 - the median price of items decreases as the condition of item worsens. This helps us explain why price of items being paid by shipping is slightly higher than those paid by seller
- Since there is a natural order in the way shipping impacts the price - we have decided to **encode it ordinally**

Analyzing item_name and item_description



- First, `name`, `updated_brand_name`, `item_description` are concatenated to generate a comprehensive textual description.
- Pretrained **BERT** encoder is then used to extract semantic information from the text, generating **768-dimensional embedding vector**
- **t-sne algorithm** is finally utilized to project these embeddings to 2-dimensional space and visualize the result.
- We can see that some small clusters(with the same color) occurring in the graph, which indicates that these items with similar descriptions tend to have close price range.

- **Sampling:** Our dataset has ~1.5 million observations in total. Hence, we can afford to split our data in the ratio **7:2:1 (train:val:test)**. Further, we plan to use **KFold cross validation for hyperparameter tuning** for the various models we plan to use.
- As suggested in the project proposal, we are dealing with a supervised regression problem here and hence the idea will be to -
 - **Start with baseline regression models** such as linear regression and its variants
 - Move on to more complex non-linear models such as **RandomForest, and XGBoost,**
 - Next we will **explore deep learning models**
 - Ultimately, we will try model ensemble to generate the best result.
- We have already **utilized BERT model to create word embeddings** for our main textual variable (item_description) and depending upon how our models perform, we might want to refine the embeddings further
 - Depending upon how our models perform, we might even want to try out different representations (like Word2Vec, or even simple TF-IDF)