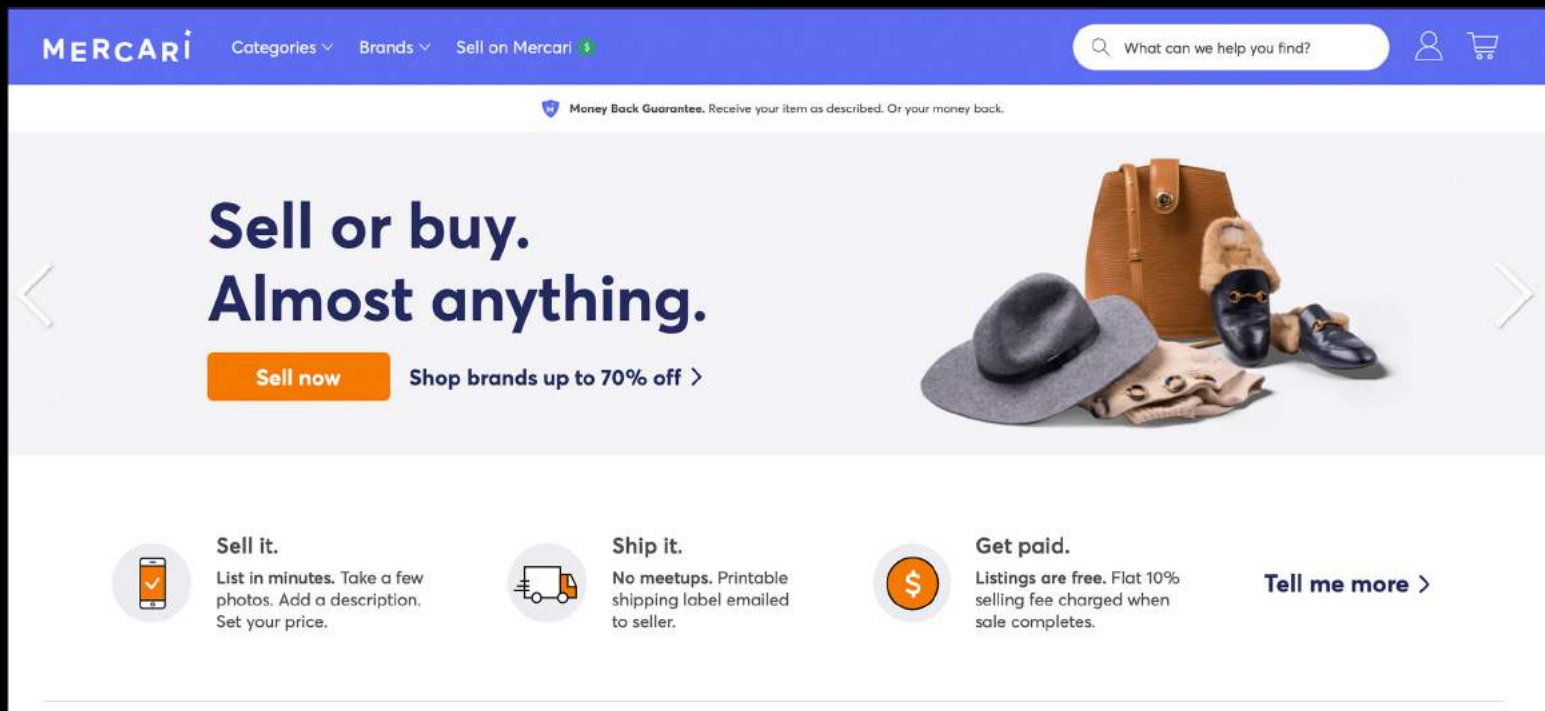


# Mercari Price Analysis Project



Mercari is an e-commerce company currently operating in Japan and the United States. Their main product, called Mercari, is a marketplace app - which has grown to become Japan's largest community-powered marketplace with over JPY 10 billion in transactions carried out on the platform each month.

A problem for Mercari is price suggestions - Mercari would like to offer their sellers price suggestions based on what products they want to sell. However, this is tough, because sellers are enabled to put just about anything on the marketplace.

To help solve this problem, Mercari have put up a public dataset concerning products that have been sold on their marketplace. In this dataset, each product sold has associated properties like the name of the listing, the item description of product, the brand of the product, and more. The idea is to build an algorithm, with the dataset, that can offer price suggestions of products sellers wants to put up on their marketplace platform.

The challenge was originally posted on Kaggle:

<https://www.kaggle.com/c/mercari-price-suggestion-challenge/overview>

In this paper, the key insights of a data analysis approach to the problem is presented, and the key findings in building a linear model and applying the model to the dataset is presented.

The results indicate that the model can suggest prices for products, generally, really well in most cases, but with some difficulty suggesting prices for products that historically have been sold for very high prices

The underlying code can be found on:

<https://github.com/wildanwildan94/Mercari-Price-Analysis---Inference-Prediction-of-Products>

# Mercari Price Analysis - Data Analysis - Description of dataset, item condition of products

- For an idea of the dataset, we consider the different attributes that are available and some typical associated values

## · Generic Values of Each Attribute:

Name: Smashbox primer

Item Condition Id: 2

Category Name: Beauty/Makeup/Face

Brand Name: Tarte

Price: 8.0

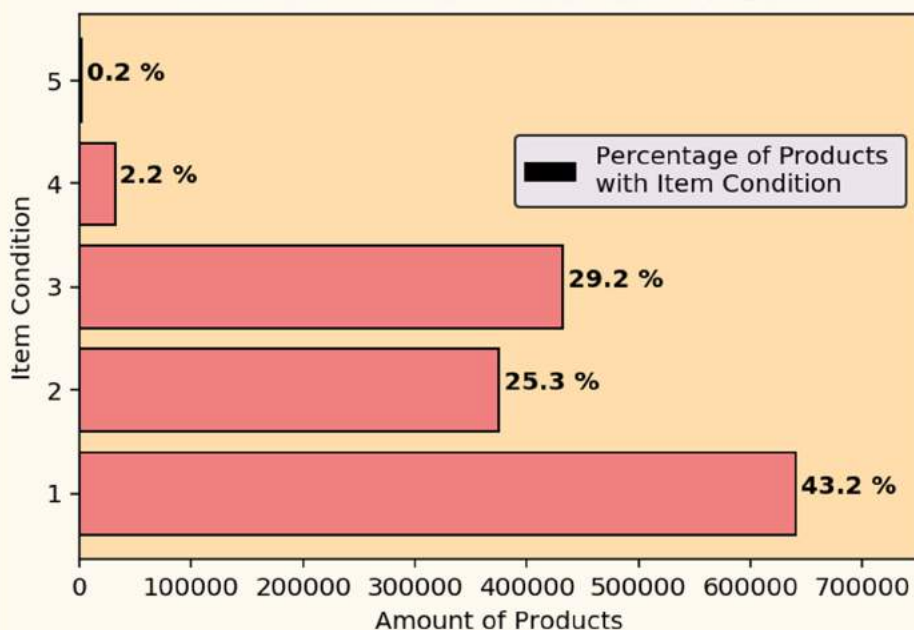
Shipping: 1

Item Description: 0.25 oz Full size is 1oz for [rm] in Sephora

- The Name is a typical, brief description of the the product in question
- The Item Condition is a number representing the condition of the product in question
- The Category Name represents the category of the product
- The Brand Name is simply the brand of the underlying product, e.g. Nike
- The Price is the price the product as sold for, in the unit USD
- The Shipping is 1 if the shipping fee is paid by the seller, and 0 if it is paid by the buyer

- For each product there is an associated item condition describing the quality of the product - let us look at that

Amount of Products in each Item Condition



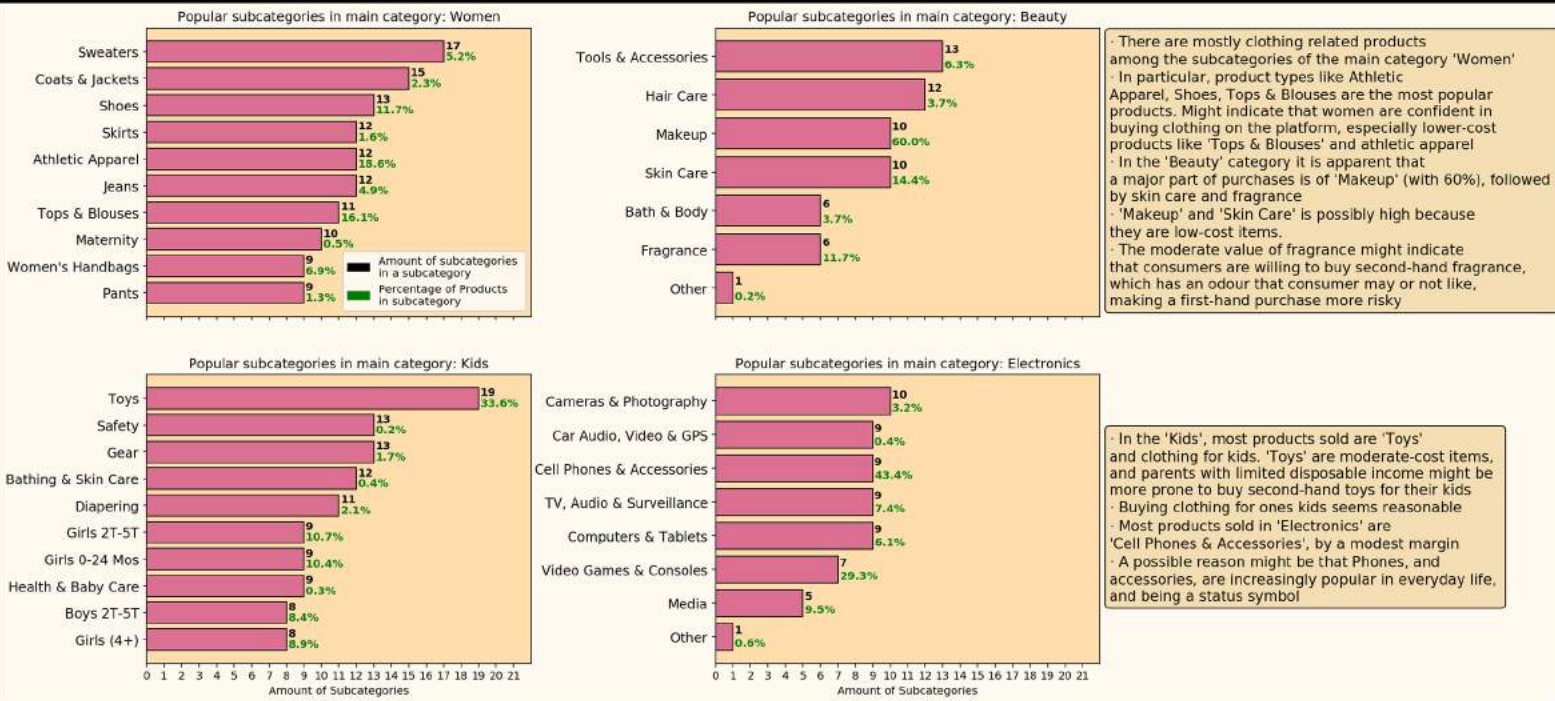
- 1: New
- 2: Almost New
- 3: Good
- 4: Fair
- 5: Poor

- Most products are New, followed by products in Good condition and products in Almost New condition
- A low percentage of products are either Fair or Poor, an indication that most people don't bother to post products in bad conditions
- A possible explanation is that most people tend to sell recently bought item, by e.g. regret or some other reason
- It may also indicate that buyers are mostly interested in products that are relatively new, and generally don't bother buying products with a low condition, because of e.g. less of a status symbol having low condition products



# Mercari Price Analysis - Data Analysis - Subcategories in certain main categories, and a description of categories

- It is of interest to analyze what type of products exists in each main category, for this we consider some popular subcategories inside a few main categories



-The ubiquity in the amount of categories begs the question of how many categories exists at each depth

**Want to quantitatively analyze the depth of categories**  
**- Is all subcategories for a product necessary?**

**The amount of categories with a certain depth:**

Depth	Amount of Products	Amount of Categories
3	1471819	1280
4	1330	5
5	3059	2
Total	1476208	1287

**The categories with a depth of 4:**

- Handmade/Housewares/Entertaining/Serving
- Men/Coats & Jackets/Flight/Bomber
- Men/Coats & Jackets/Varsity/Baseball
- Sports & Outdoors/Exercise/Dance/Ballet
- Sports & Outdoors/Outdoors/Indoor/Outdoor Games

**The categories with a depth of 5:**

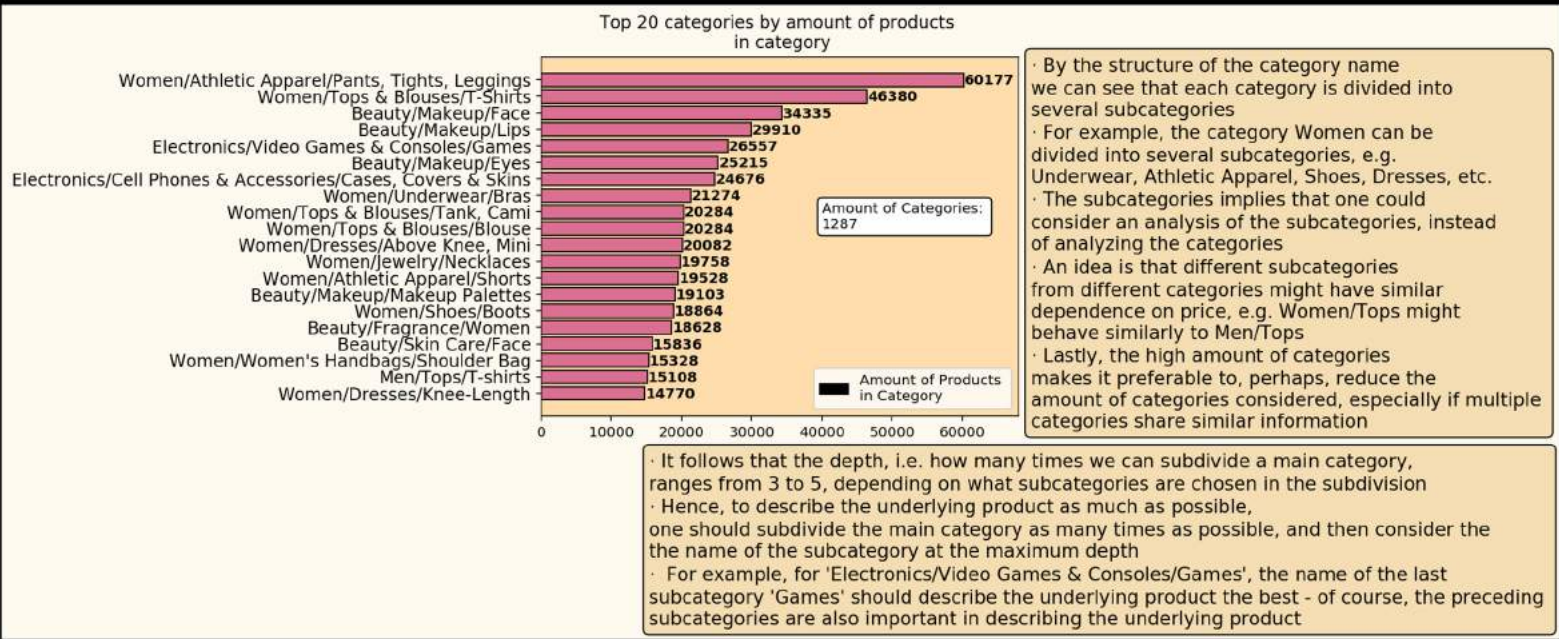
- Electronics/Computers & Tablets/iPad/Tablet/eBook Access
- Electronics/Computers & Tablets/iPad/Tablet/eBook Readers

**From the structure and low quantity of the categories with a depth of 4 and 5, we can reconsider the categories as:**

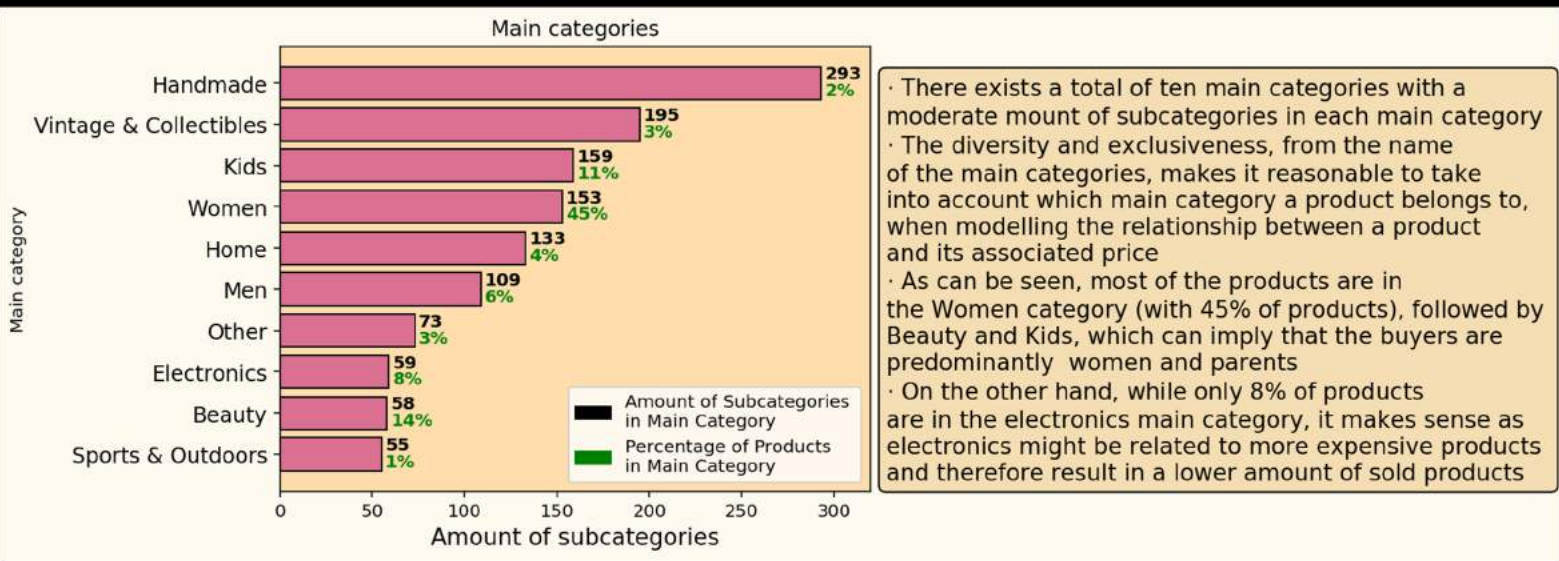
- Handmade/Housewares/Entertaining Serving
- Men/Coats & Jackets/Flight Bomber
- Men/Coats & Jackets /Varsity Baseball
- Sports & Outdoors/Exercise/Dance Ballet
- Sports & Outdoors/Outdoors/Indoor Outdoor Games
- Electronics/Computers & Tablets/iPad Tablet eBook Access
- Electronics/Computers & Tablets/iPad Tablet eBook Readers

# Mercari Price Analysis - Data Analysis - Categories and the main categories of the dataset

- For an idea of what type of categories each product belongs to, we consider an analyze of the category attribute -What categories exists? What are their names?



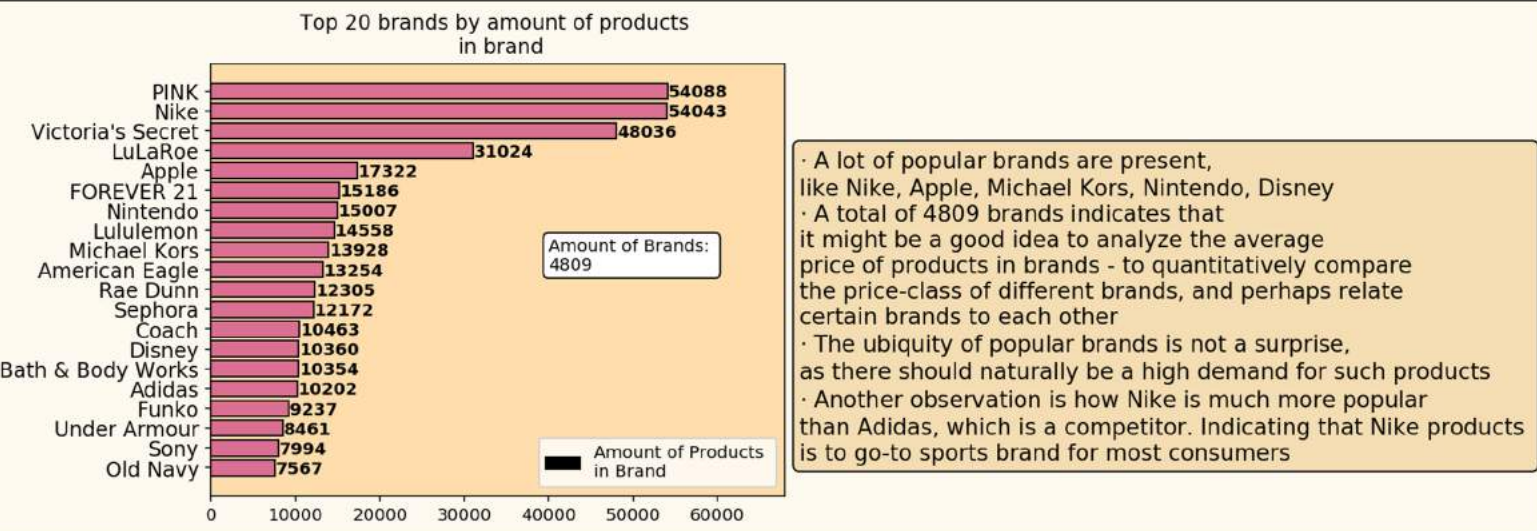
- To continue, we consider the different main categories that exists, to get a rough idea of what type of products exists





# Mercari Price Analysis - Data Analysis - Brands of products, and the price of products

- Another important property of each product is what brand it belongs to - if any - as the brand would most likely have an effect on the price



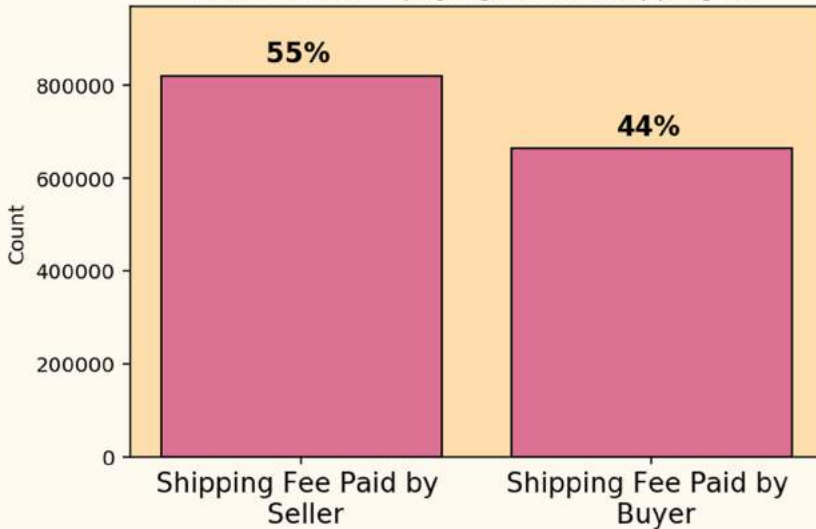
- The attribute of most interest is the price attribute, which is the price of the underlying product.



# Mercari Price Analysis - Data Analysis - Shipping fee and the name of each product

- For each product bought and sold on the platform, there is an associated shipping fee that is either paid by the buyer or the seller - what is its distribution?

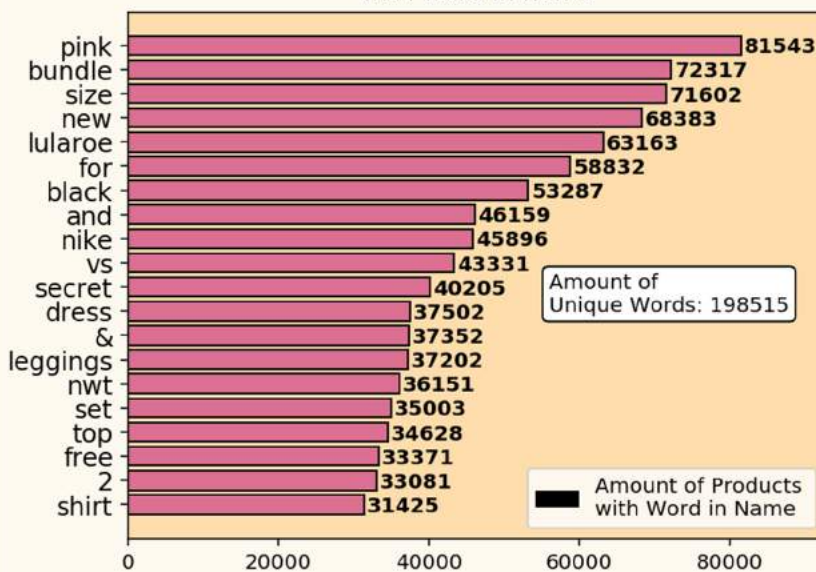
Count of who is paying for the shipping fee



- The shipping fee is usually paid by the seller, but not by a wide margin
- The higher percentage for the shipping fee paid by the seller can indicate that there is a higher supply than demand, such that the sellers are forced to buy shipping fee more often to, perhaps, make their product listings more attractive to potential buyers
- Another possibility is that the seller paying for shipping might be more attractive to potential buyers - it signals a burden (both practical and economical) left to the seller

- Each product has an associated name, which is used to present the product listing to the potential buyer. Potentially, the words used in the name can be of interest in modelling the product's underlying price

Top 20 words by amount of products with word in name

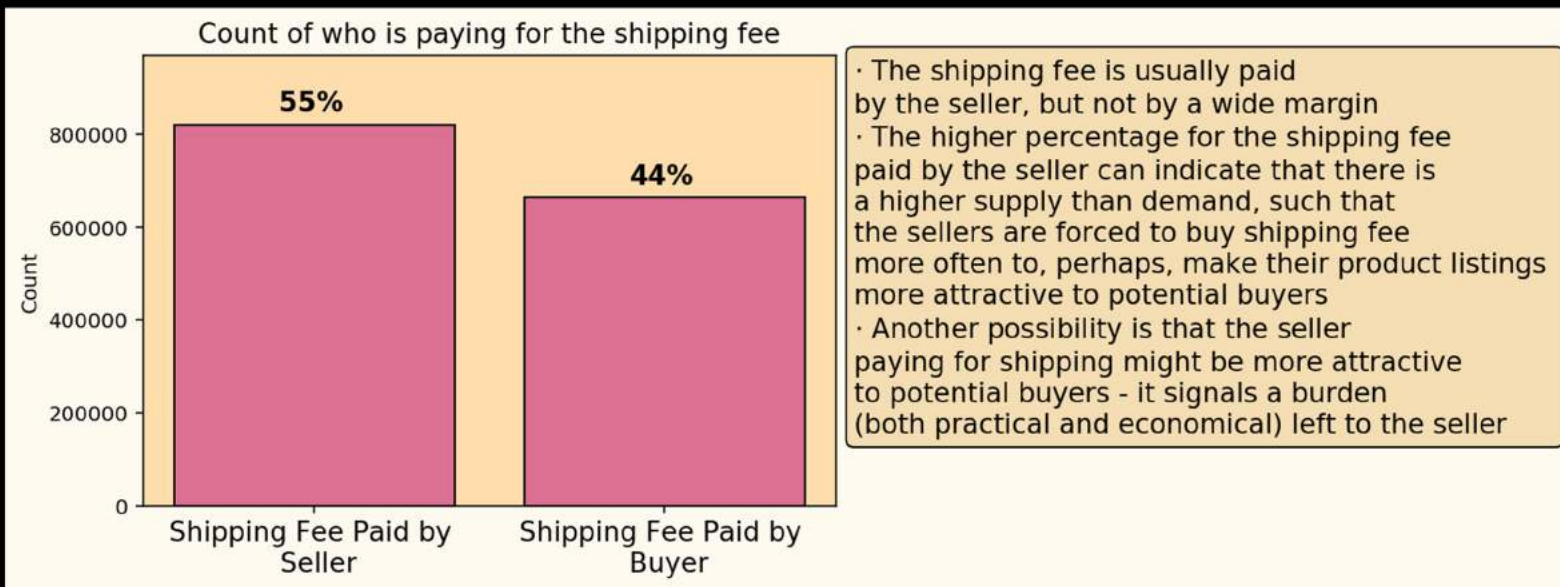


- A lot of descriptive words in the top 20 words in the name attribute
- For example pink, bundle, black, dress, leggings, top, shirt
- Indication that the words in the name attribute can describe the underlying product (e.g pink shirt)
- There exists a few non-descriptive words like vs, &, free, but they are relatively few
- A conclusion is that the words in the name could potentially be used, in some way, to model what type of product is underlying the name
- Drawbacks include the large amount of unique words, which is at 198515, and that the name attribute contains a lot of special characters, which can distort the meaning of a word

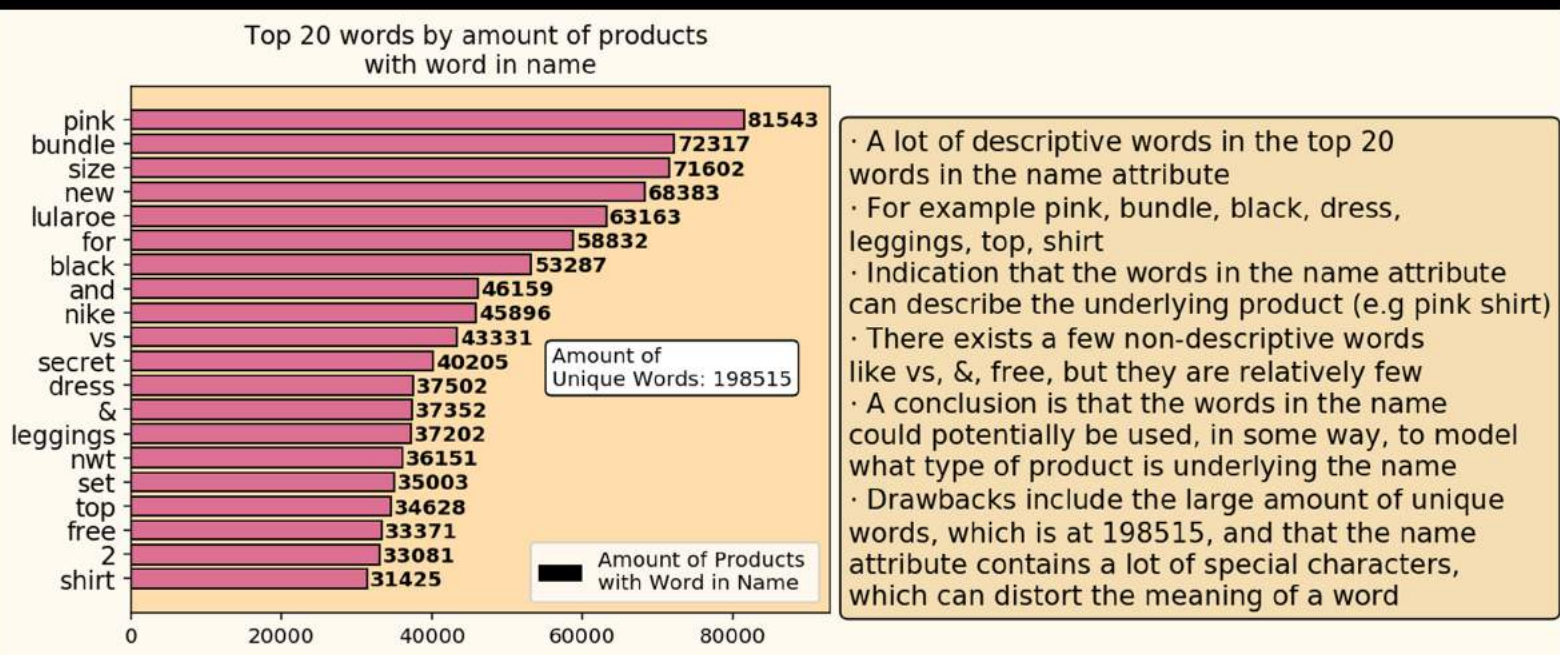


# Mercari Price Analysis - Data Analysis - Shipping fee and the name of each product

- For each product bought and sold on the platform, there is an associated shipping fee that is either paid by the buyer or the seller - what is its distribution?



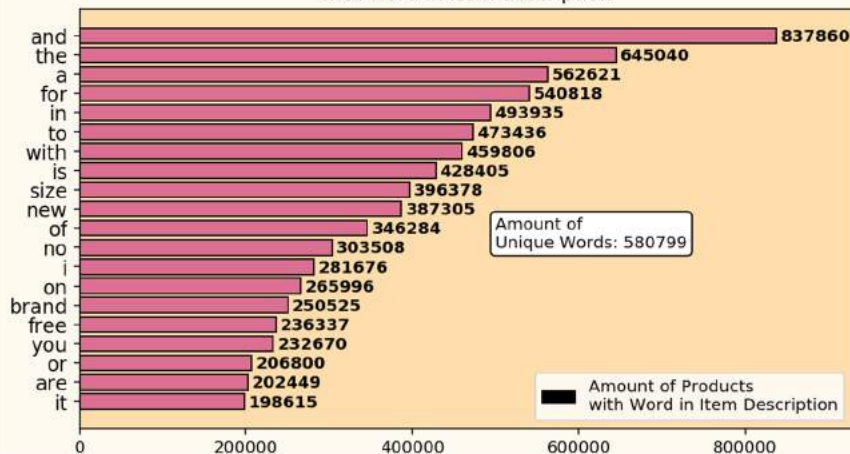
- Each product has an associated name, which is used to present the product listing to the potential buyer. Potentially, the words used in the name can be of interest in modelling the product's underlying price



# Mercari Price - Data Analysis - Words in item description and the average price associated with words in the name

- Similarly as in the name case, the item description might contain a lot of keywords that can describe the underlying product in the listing well

Top 20 words by amount of products with word in item description



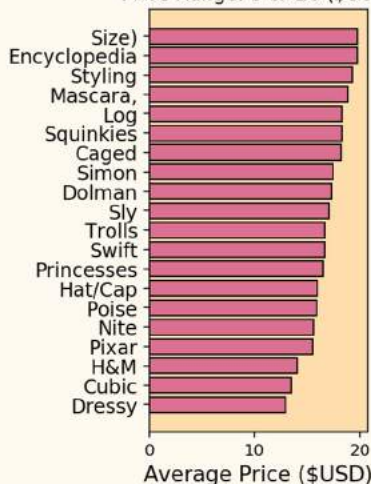
- A lot of non-descriptive words in the top 20 words in the item description
- Might imply that utilizing words in the item description might not be useful, as it contains a lot of sentence-building words, like and, the, size, brand, free, on
- In addition, a lot of words contain special characters which might distort the meaning of words
- However, all the words in the item description might collectively convey useful information, for example if the words 'size' appears with '8', it might indicate that the underlying product is a clothing piece with size 8. This type of inference should be able to convey some information on the possible price of the underlying item
- Hence, the words collectively might convey a lot of useful information

- In addition to what words exists in the name of products, there is an interest in correlating those words with prices of underlying products

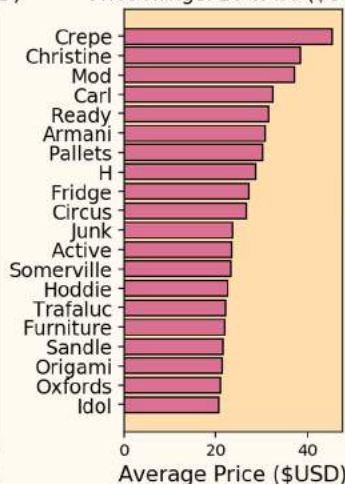
Average of Prices Related to Word in Name

20 Randomly Picked Words for Each Price Range

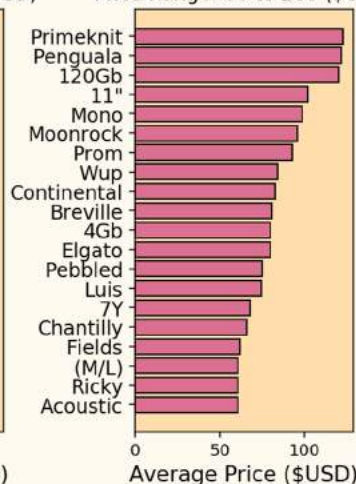
Price Range: 5 to 20 (\$USD)



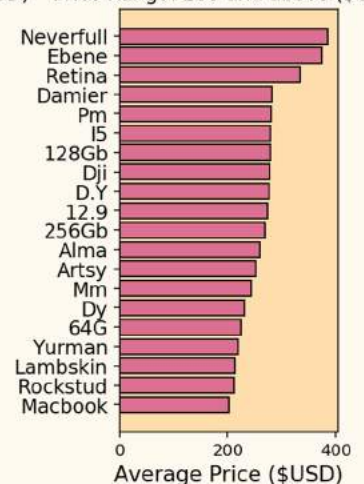
Price Range: 20 to 60 (\$USD)



Price Range: 60 to 200 (\$USD)



Price Range: 200 and above (\$USD)



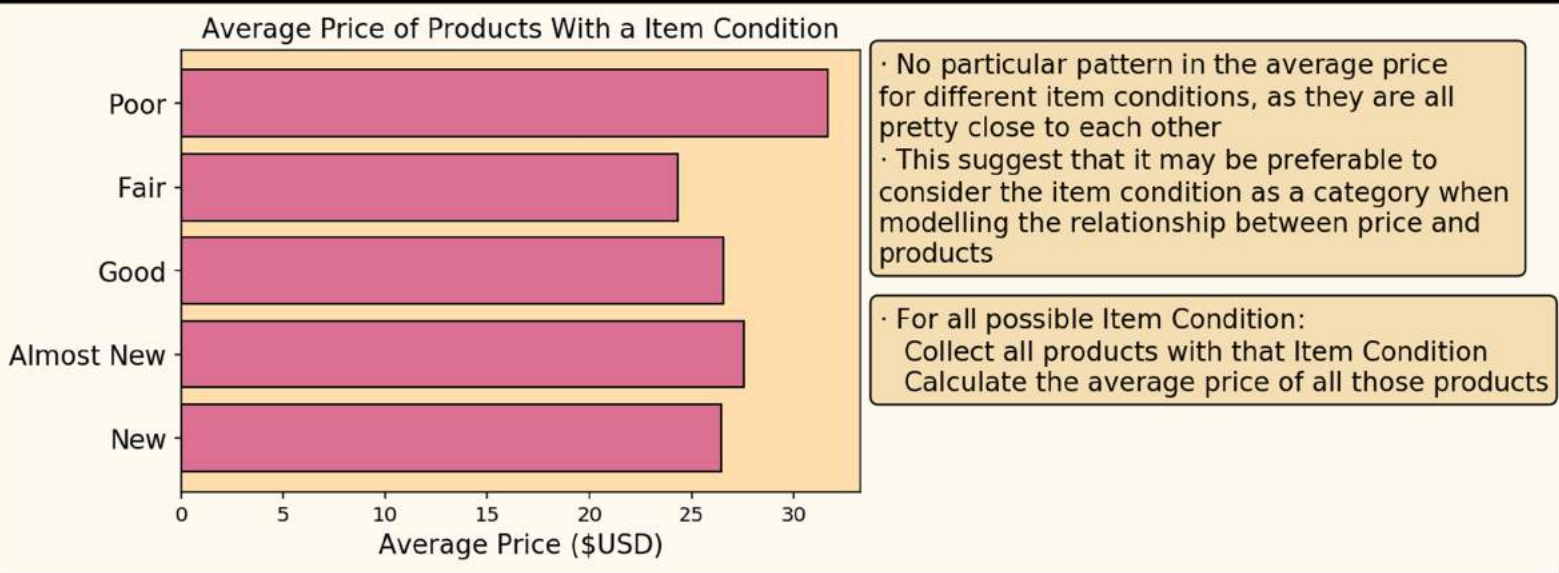
- A lot of descriptive words in all price ranges, including words like Trolls, Princesses, Armani, Hoddie, Oxfords, Acoustic, 120Gb, 256Gb, Macbook, Lambskin, Damier
- Some words give an almost full description of the underlying product, e.g. Macbook, while some words give an important aspect, e.g. Lambskin
- Hence, it is concluded that the words, especially together, in the name attribute of a product can give potentially useful information of a product's underlying price

- Construction of the Average Price:  
For each possible Word in the name attribute:  
Collect all products which has that Word in its name attribute  
Calculate the average price from all the prices of those products

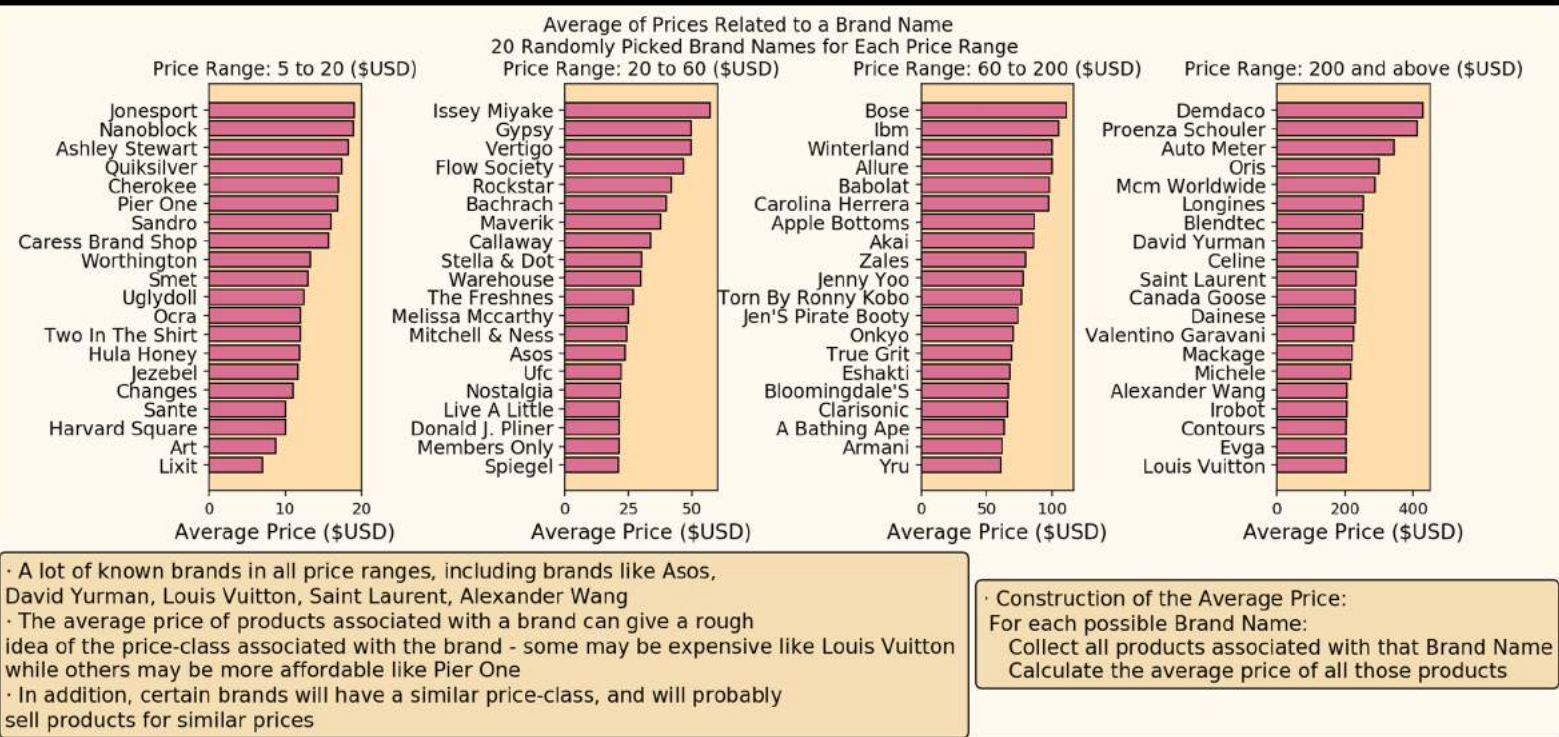


# Mercari Price Analysis - Data Analysis - Prices associated with item condition of products, and prices of products in different brands

- A minor interest is to consider the prices of products with different item conditions - are less quality items generally cheaper?

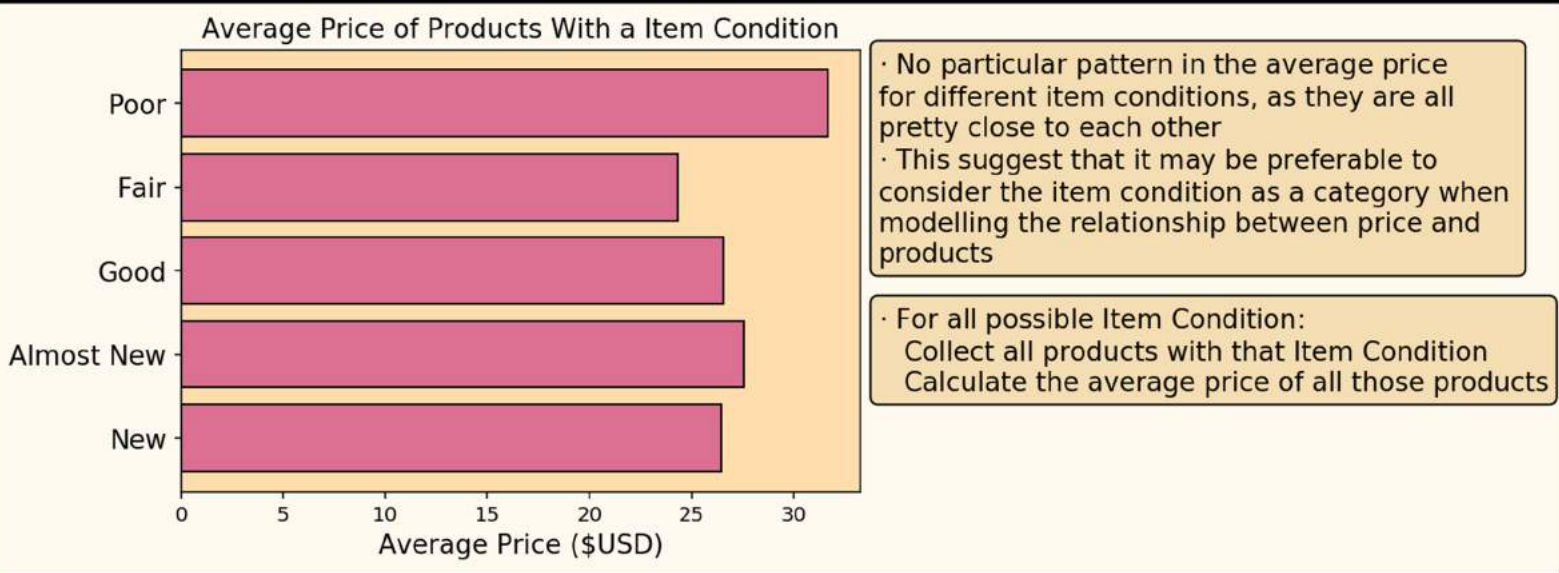


- A major factor of the price will obviously be which brand the underlying product belongs to - especially what type of price products in that brand usually sell for, in a general sense

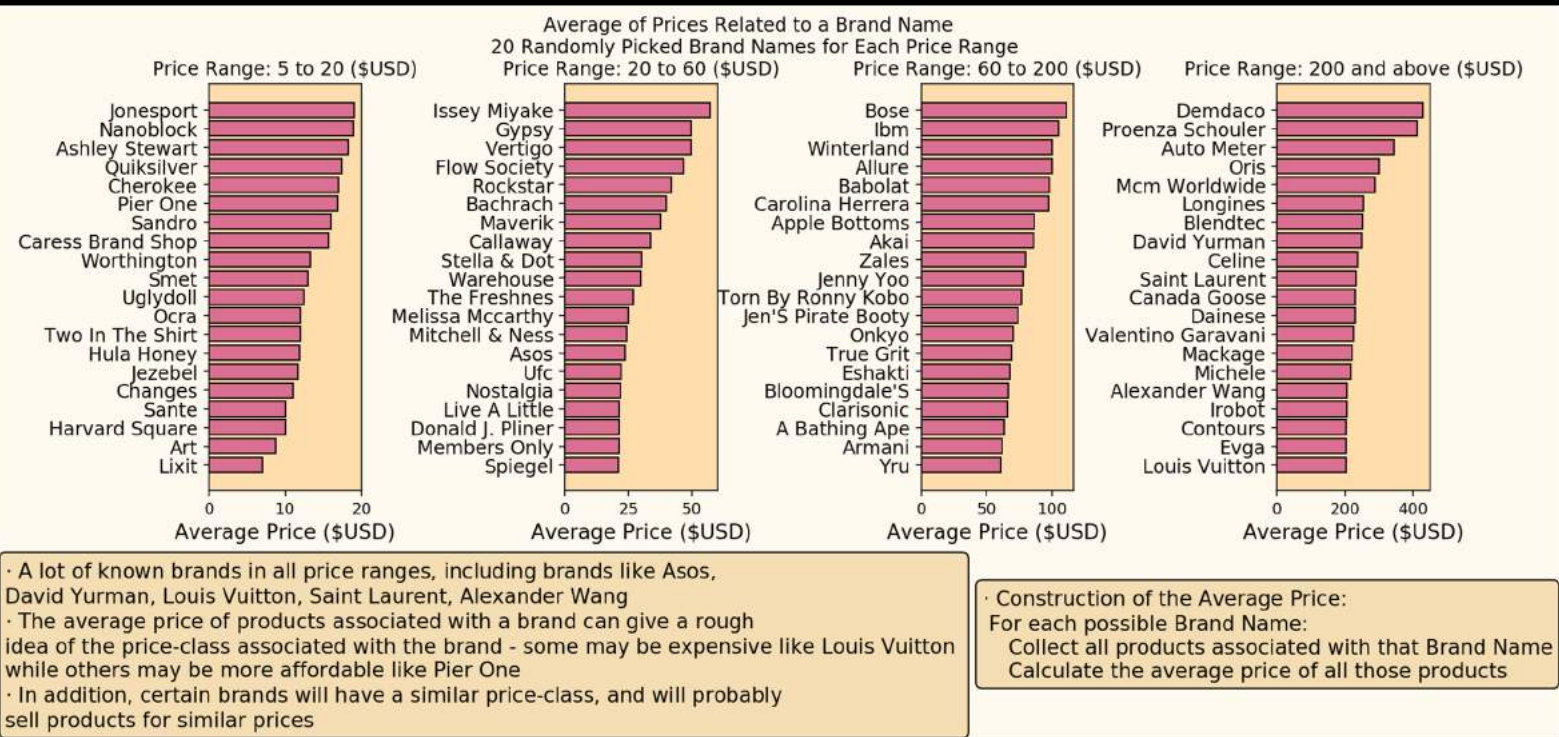


# Mercari Price Analysis - Data Analysis - Prices associated with item condition of products, and prices of products in different brands

- A minor interest is to consider the prices of products with different item conditions - are less quality items generally cheaper?



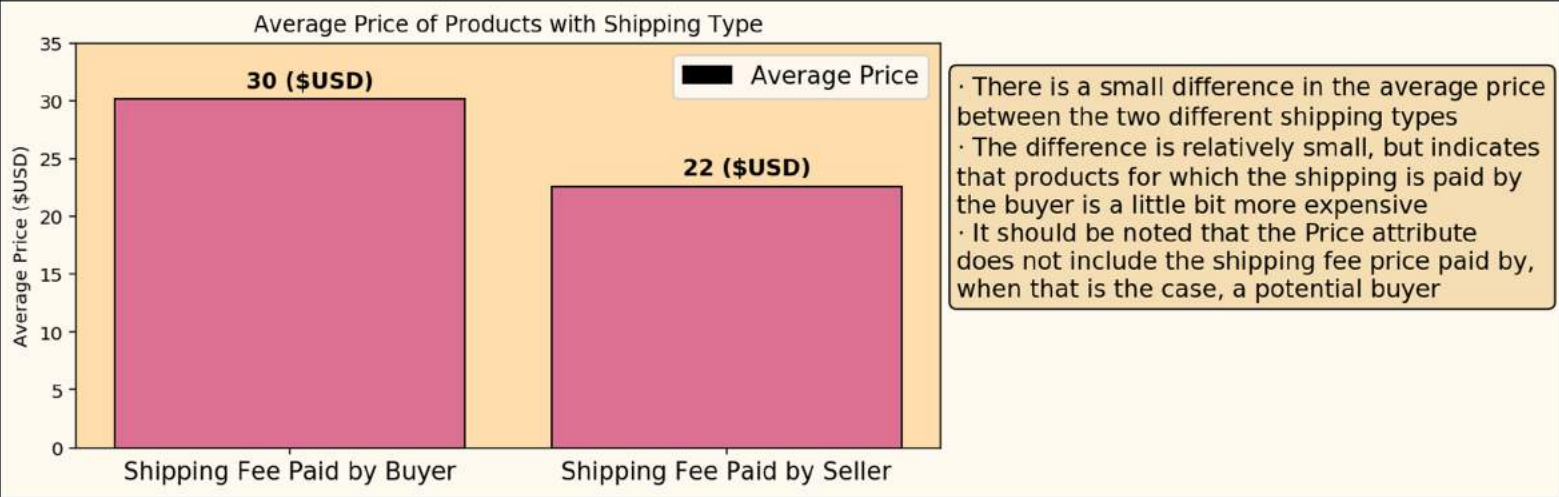
- A major factor of the price will obviously be which brand the underlying product belongs to - especially what type of price products in that brand usually sell for, in a general sense



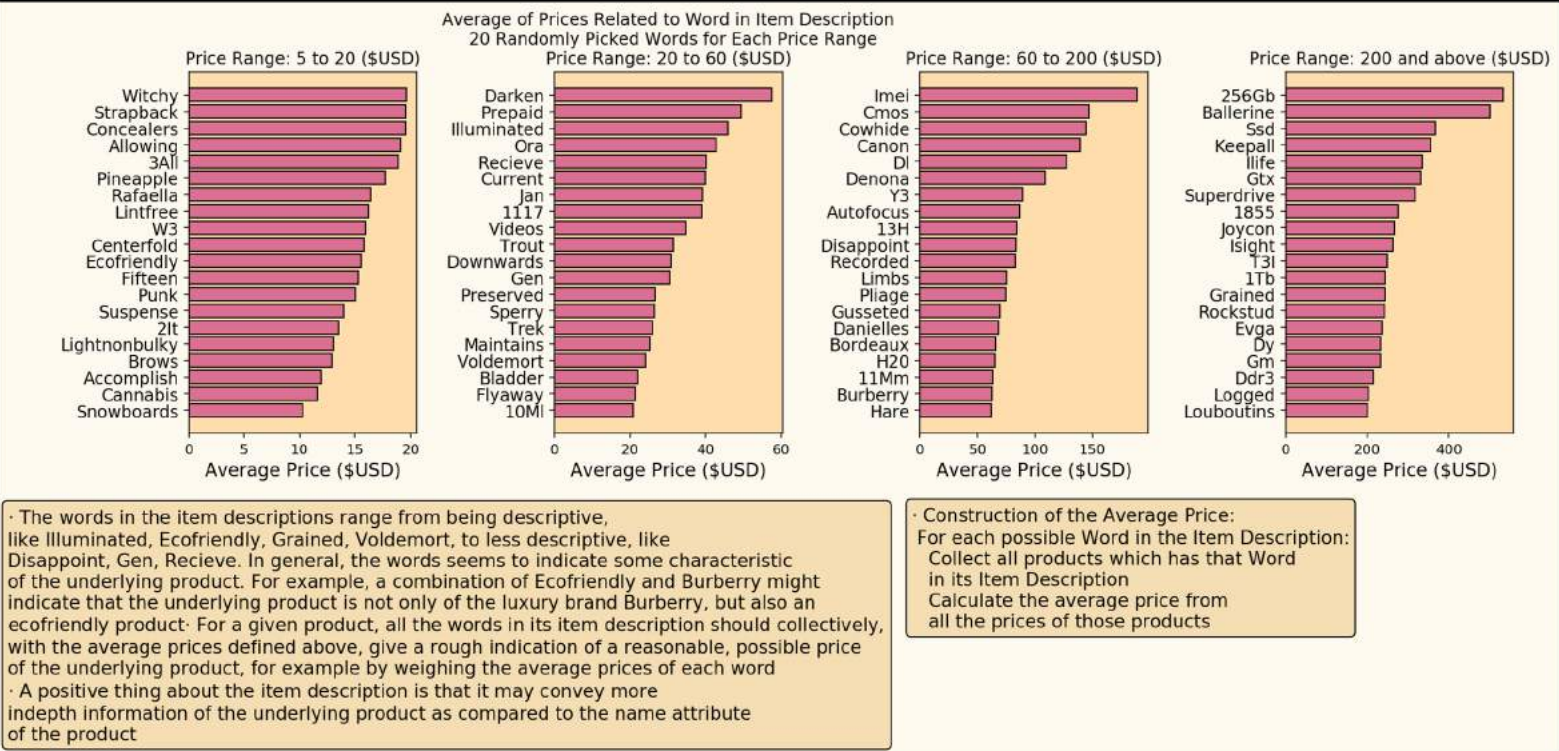


# Mercari Price Analysis - Data Analysis - Shipping fee with price, and the price associated with words in the item description

- Further considering the shipping fee of items, there might be a relation between who pays the shipping fee and the average price of products



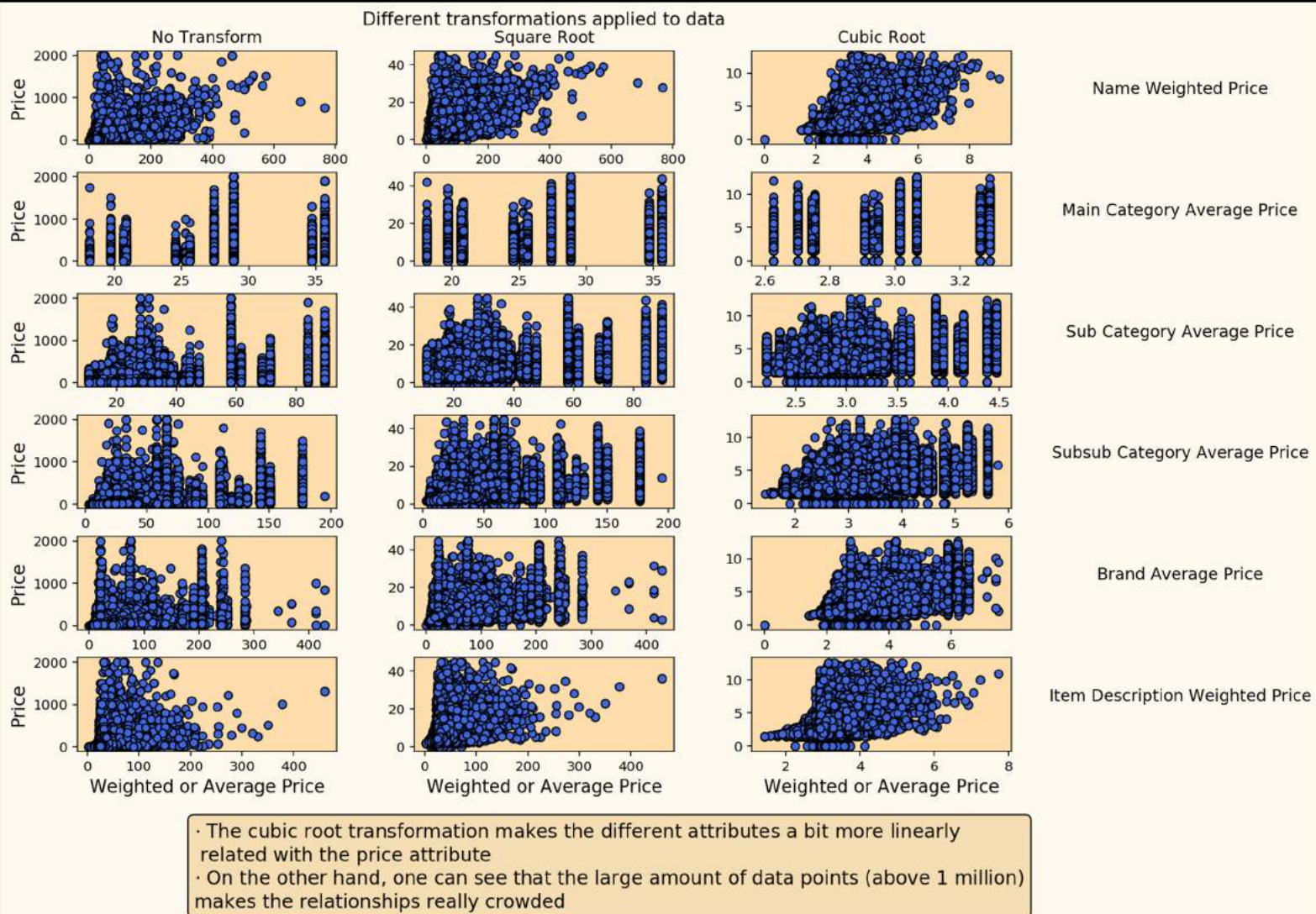
- The item description might convey important information of the underlying product - hence, the price associated with words in the item description is of interest



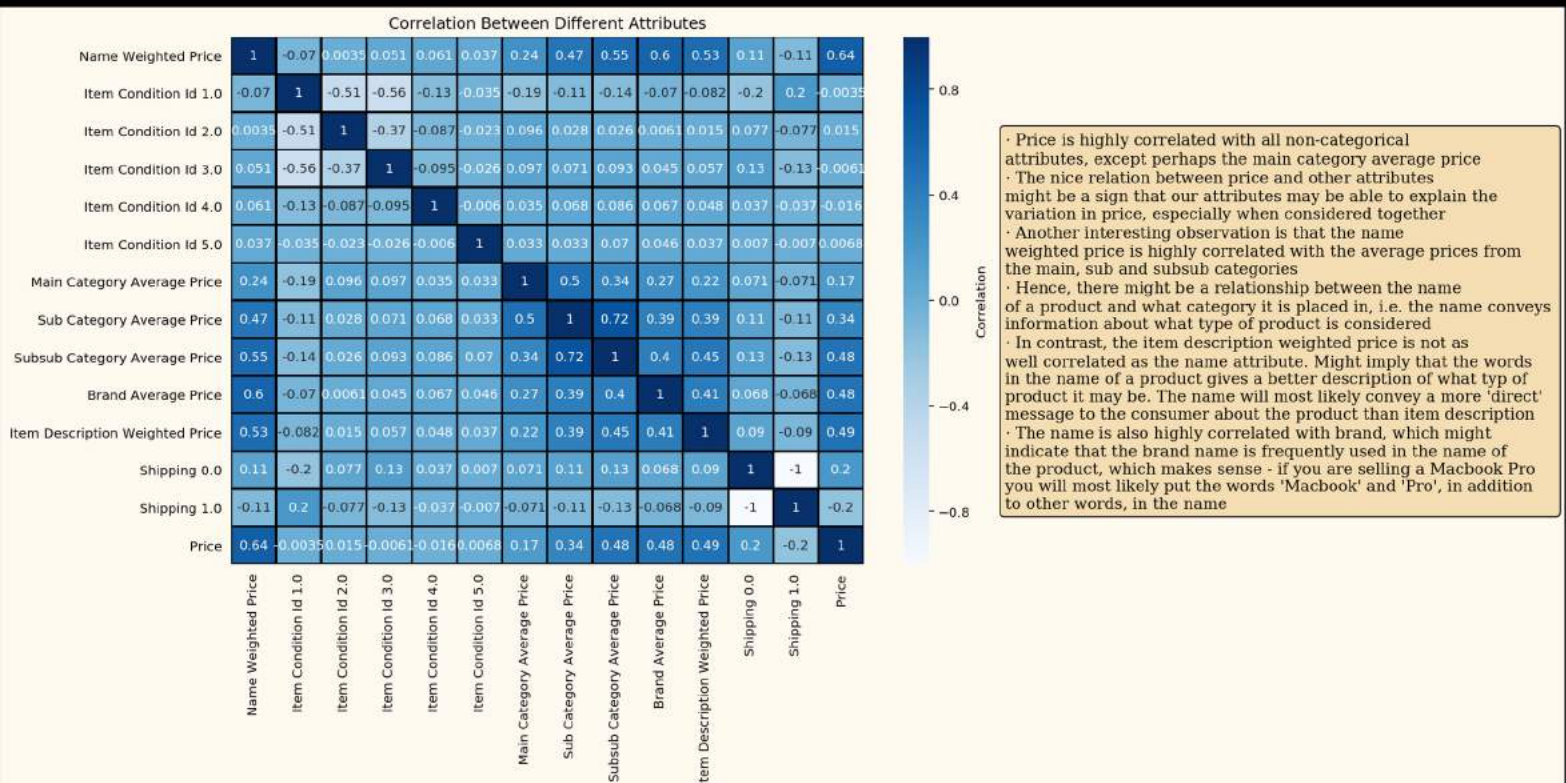


# Mercari Price Analysis - Prediction Analysis - Transformation of variables, correlation among variables

- Before building a linear model, there might be some transformations of variables that may help with relationships



- A key idea in linear models is to analyze the correlation among all variables considered, especially the correlations with respect to the price variable



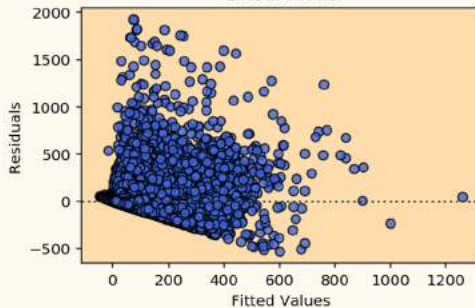


# Mercari Price Analysis - Prediction Analysis - First linear model; Diagnostics and Cook's Distance

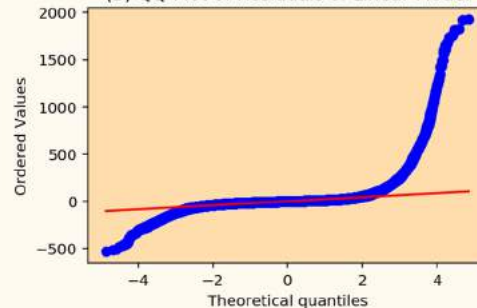
- A first linear model is applied to the problem, with the price attribute used as a response variable. To analyze the appropriateness of the model, the idea is to consider diagnostic plots

Fitted Model; A square root transformation and centering of attributes around their means are applied

(A) Fitted Values vs. Residuals of Linear Model



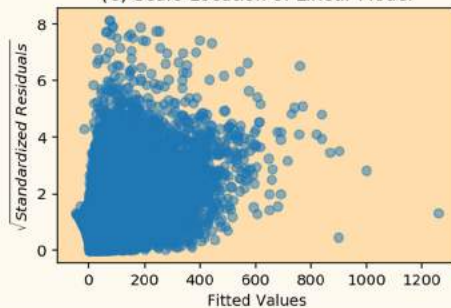
(B) QQ-Plot of Residuals of Linear Model



- (A) The spread around the horizontal line makes nonlinear residuals probable, but not perfect
- (A) In particular, the model has a weakness in predicting high price products in some cases, especially for low fitted values

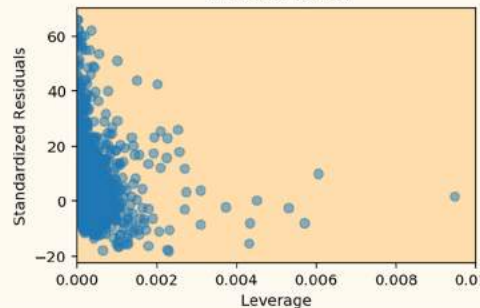
- (B) In the QQ-plot, the assumption of normally distributed residuals is somewhat probable, in the middle region, but the extreme deviations at the tails are unfortunately also clear
- (B) The extreme values indicates that the data may have too many extreme values for normally distributed residuals to be true
- (B) As in (A), the extreme values are most likely cases where the fitted value is low when the real product price is high
- (B) This signifies that the model might not always be good at predicting the price of high-price products
- (B) A possible reason might be that most data values are items which are sold for low/moderate prices, and will tilt the model to become better at predicting the price of products with low to moderate prices

(C) Scale-Location of Linear Model



- (C) A fairly equal spread indicates that constant variance of residuals is reasonable

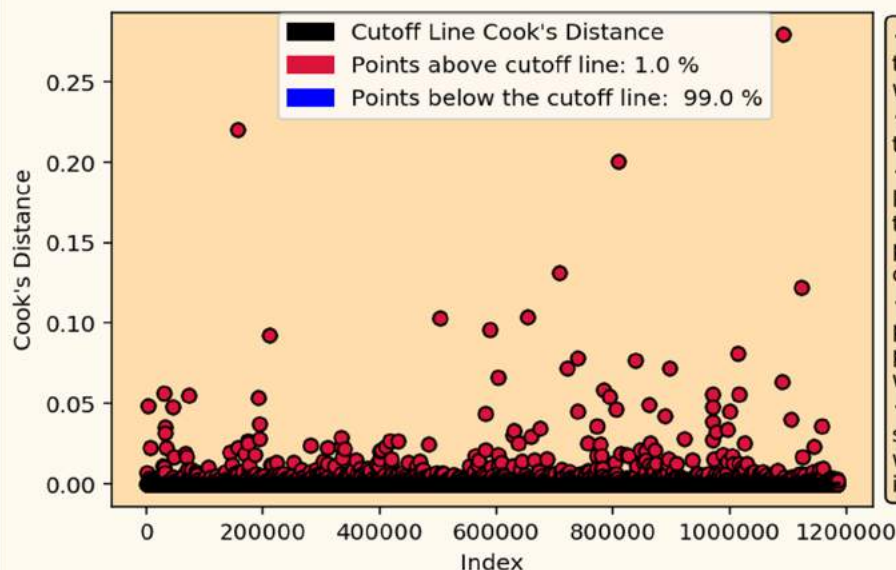
(D) Leverage vs. Standardised Residuals of Linear Model



- (D) Most values with high standardised residuals are not influential
- (D) There are a few points with high leverage, but with low to moderate standardised residuals, which makes them less influential on the fit of the linear model

- Further describing the linear model, especially illustrating data points with influence on the fit of the model, the Cook's Distance can be considered

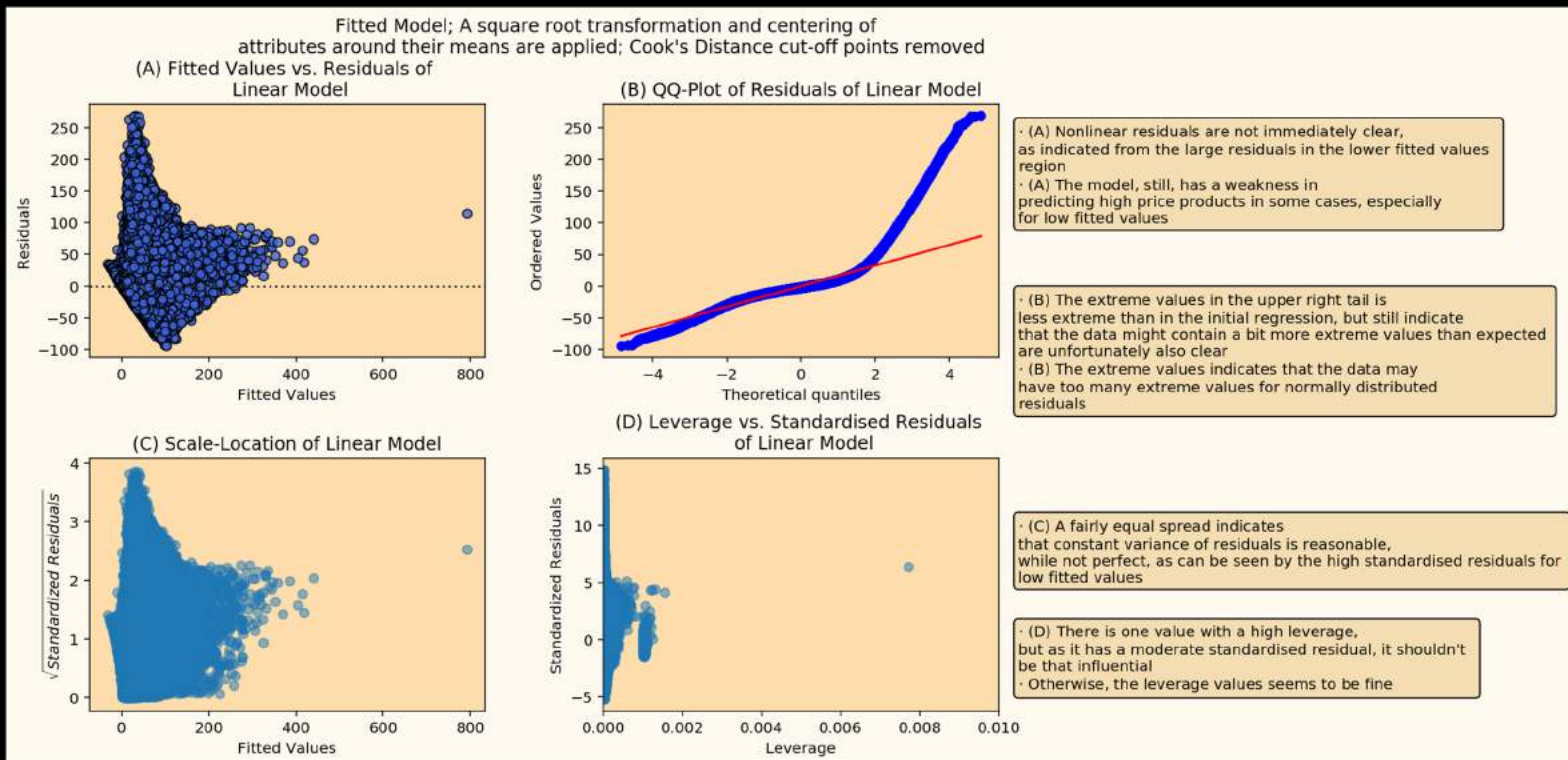
Cook's Distance of Fitted Values



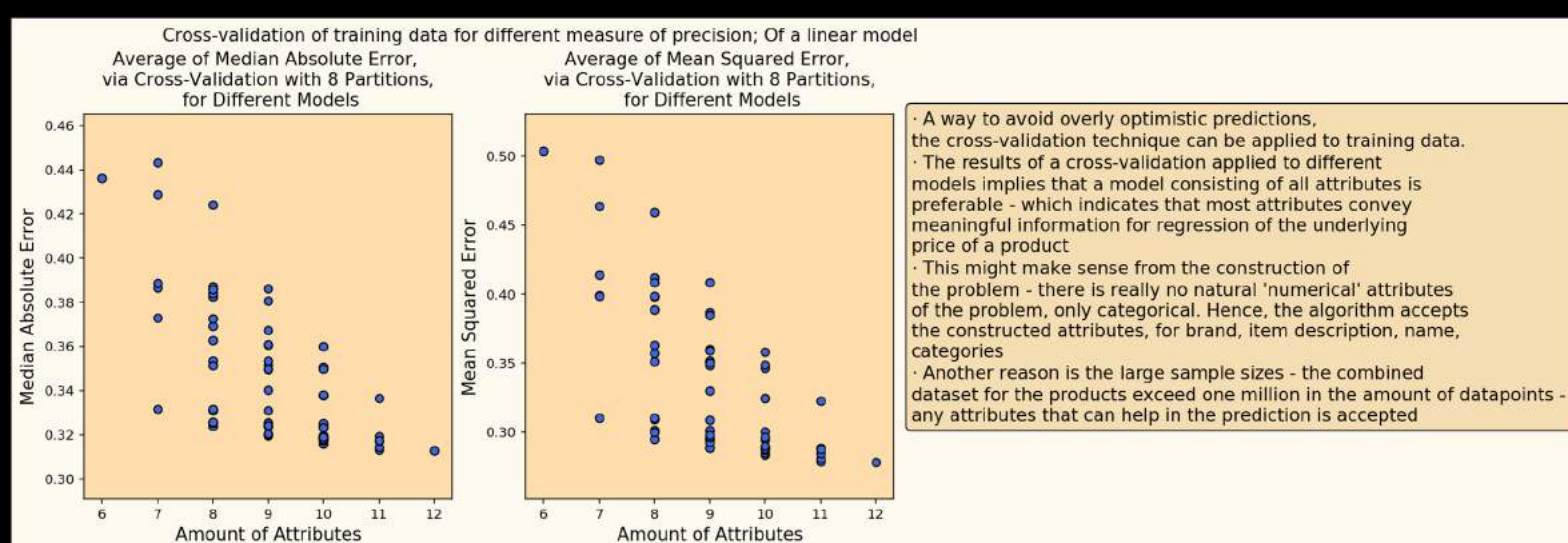
- The 1 % of points above the cutoff line indicates extreme values with great influence on the linear model
- One possible remedy is to remove these points from the linear model
- The influential points are most likely products with a high price, which may distort the prediction of price of low-moderate price products - which are more frequent in the dataset
- If we want our model to perform better for products with low-moderate prices, then the 1% of points above the cutoff line should be discarded when building the model
- Because they are only 1 %, it makes sense to remove these extreme values. However, if one wishes to retain the ability of predicting high price items, one should probably keep the 1 % of points

# Mercari Price Analysis - Prediction Analysis - Linear model after Cook's Distance analysis, and different performance measures to choose an optimal model

- After removing datapoints based on Cook's distance, another linear model is built and its properties are analyzed



- The aim is to find an optimal model. For this, we consider all combinations of attributes (where all binary attributes are kept in the model), and for each combination we evaluate two performance measures. Lastly, based on the performances, we pick a final, optimal model

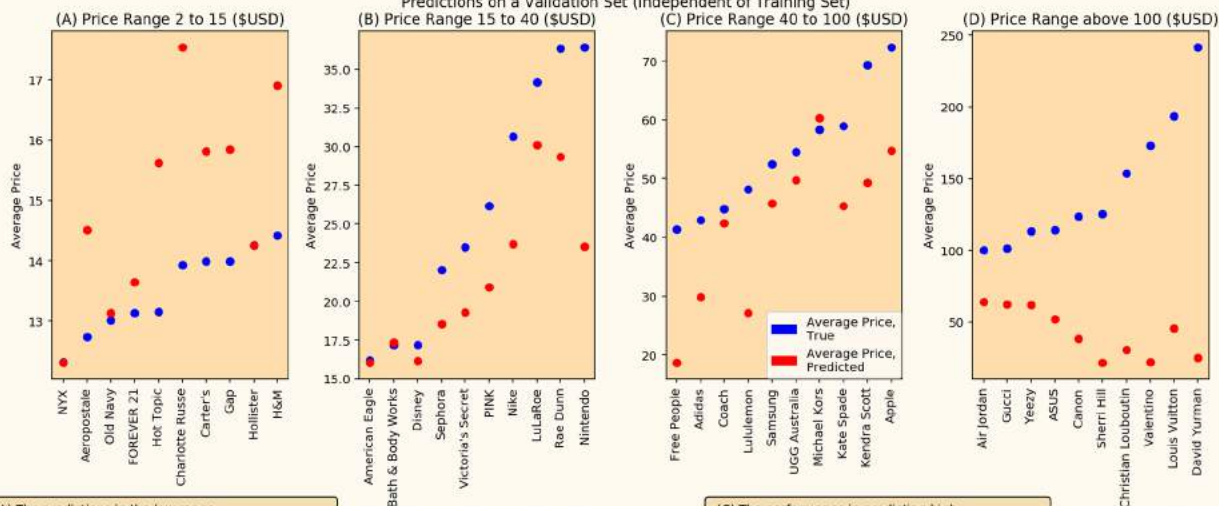




# Mercari Price Analysis - Prediction Analysis - Optimal model in predicting prices of products in brands, and predicting products in different main categories

- With our final model, the aim is to evaluate the model on datasets not used in designing the model. The idea is measure how well the model performs in practice. For a starter, we consider predictions on products in brands - to see how well the model performs when different brands of products is considered

Predicted and True Average Price of Products in Brands; Ten Brands in Various Price Ranges; Predictions on a Validation Set (Independent of Training Set)



· (A) The predictions in the low range performs remarkably well, where all predictions are just a few dollars away from the true values.  
· (A) Indicates that our model is in general good for predicting product prices for low-price range brands like H&M, Old Navy and Gap  
· (A) A possible reason might be that a lot of products exists in the low-priced range, which makes our model really efficient in predicting such products

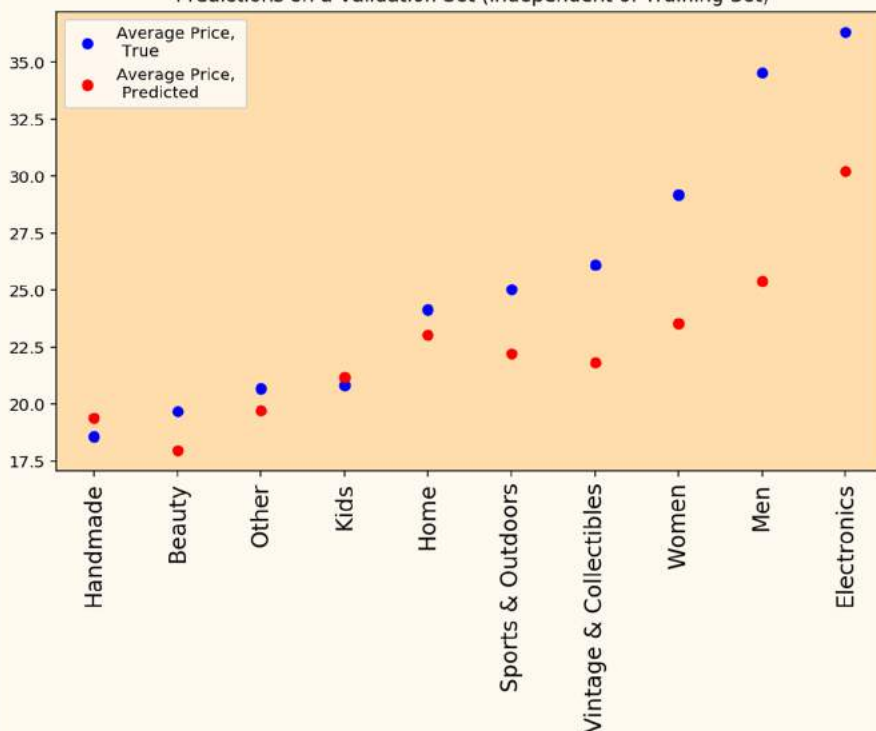
· (B) Similarly as in the low-price range, the prediction of price in moderate price range is really well too - again with just a few dollars of margin  
· (B) Hence, our model should perform well predicting prices of products like Nike, Disney and American Eagle  
· (B) The well performance might come, as in (A), from the ubiquity of moderate priced products

· (C) The performance in predicting high priced products differs among brands, as can be seen.  
· (C) While a few brands can be predicted well, like Coach, Samsung, Michael Kors, the model has a difficulty predicting e.g. Free People, Lululemon  
· (C) However, in general, the predictions are not too bad. One could argue that a margin of 10-15 (\$USD) among high-priced products isn't that much, especially for brands like Apple

· (D) Contrary to the other price ranges, the predictions for very-high priced products isn't generally well, as can be seen from the increased margin as the true price increases  
· (D) Most likely, it is due to two factors. Firstly, in our model a portion of training data points corresponding to high cook's distance points were removed, which in the process have decreased our model's ability to predict prices for very-high priced products. Secondly, there is most likely not as many high-priced products in our training data, as opposed to low- moderate-priced products, which most likely have made our model biased towards predicting the correct price of low- and moderate-priced products

- Further, there is an interest in evaluating our optimal model on products from different main categories.

Predicted and True Average Prices of Products in Main Category; Predictions on a Validation Set (Independent of Training Set)



· The predicted values seem to be quite good, in general, for most of the main categories  
· However, there seems to be some deviations when predicting products in the Women, Men and Electronics main categories. Possibly because products in these categories range from low-priced to very-high priced brands. A major portion of very-high priced brands will have a negative effect on the predictions, as discussed in the predictions of products in brands

# Mercari Price Analysis - Prediction Analysis - Optimal model in predicting products in different categories

- Lastly, there is an interest in picking a few particular categories of products and evaluate our optimal model on these categories. The idea is that these categories corresponds to a diverse set of products, and this will show how the model performs on a diverse set of categories

