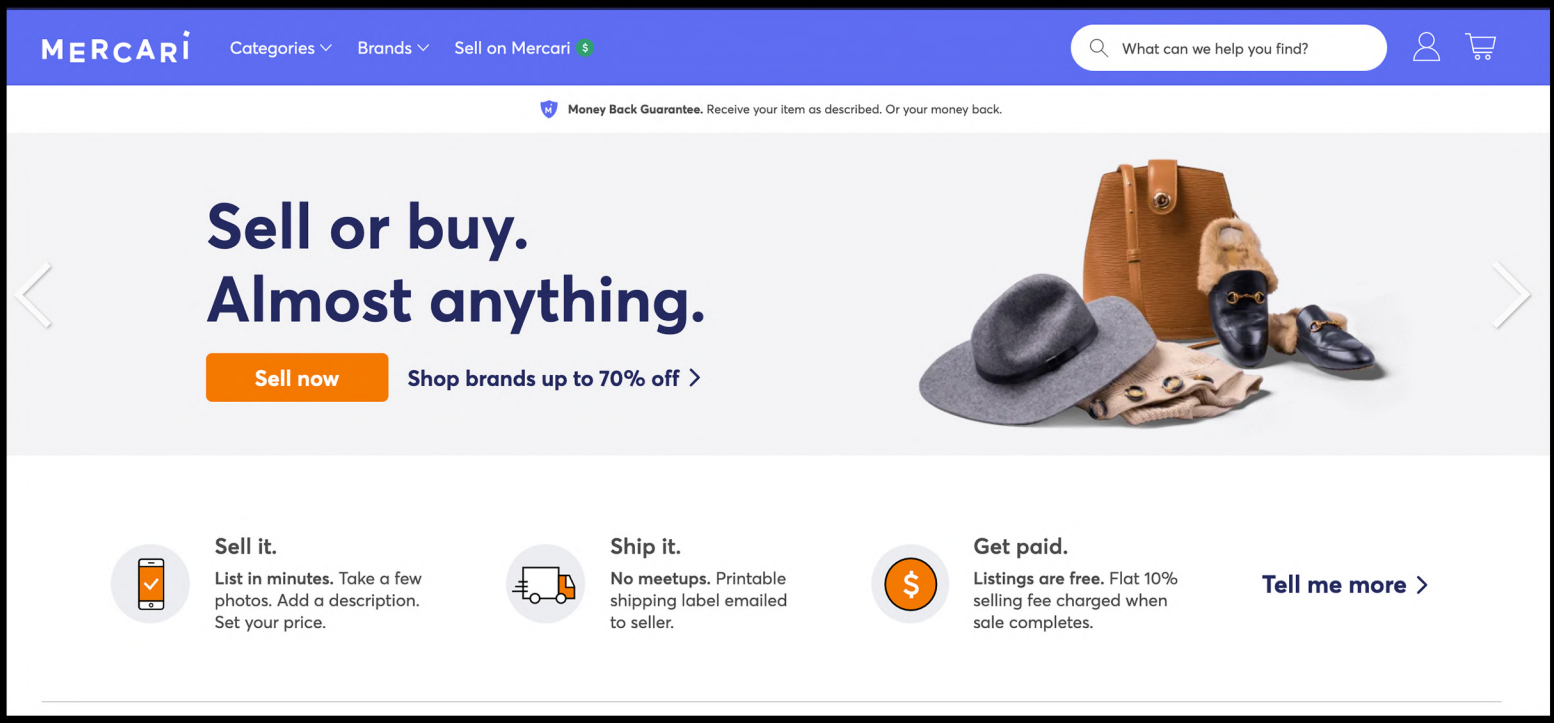
Mercari Price Analysis Project



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 markerplace.

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

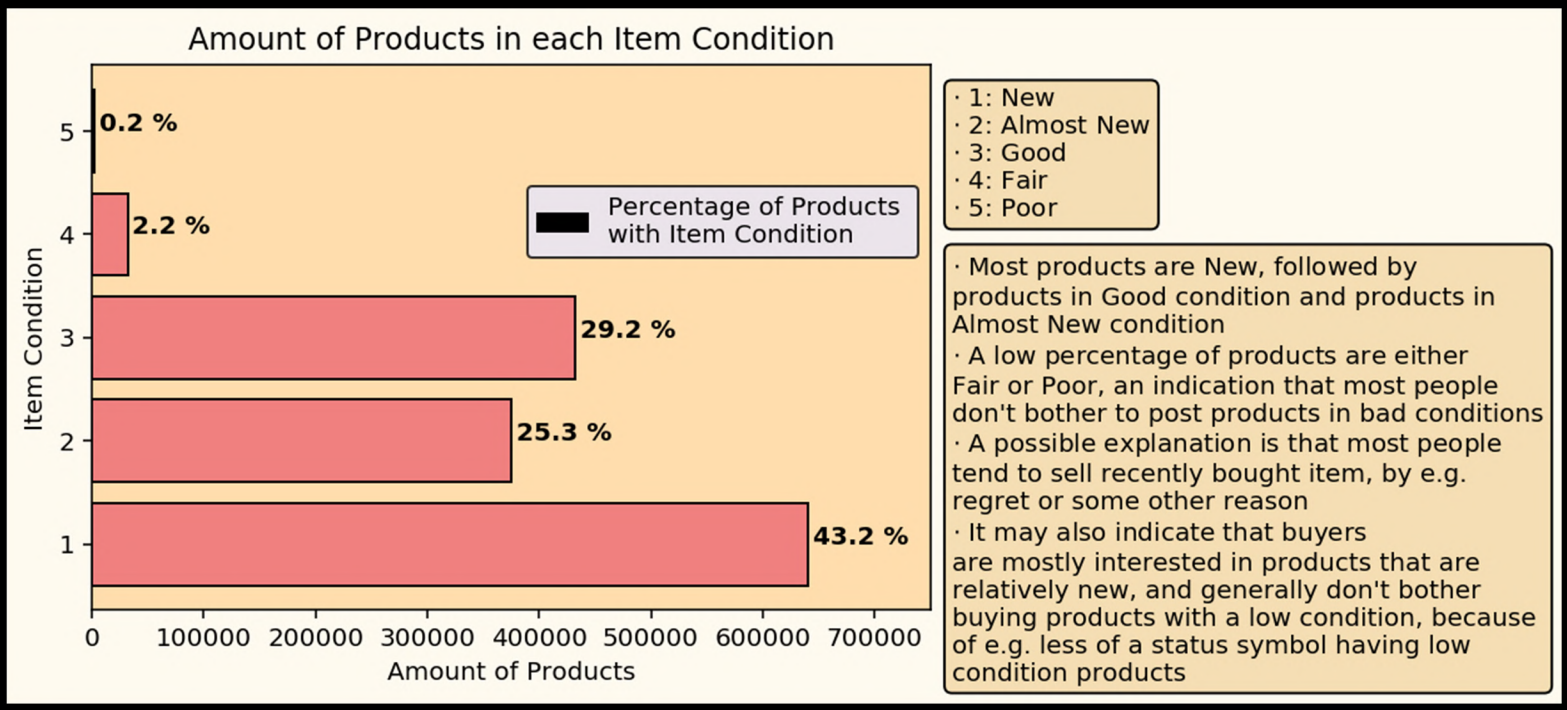
Mercari Price Analysis - Data Analysis - Description of dataset, item condition of products

- 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

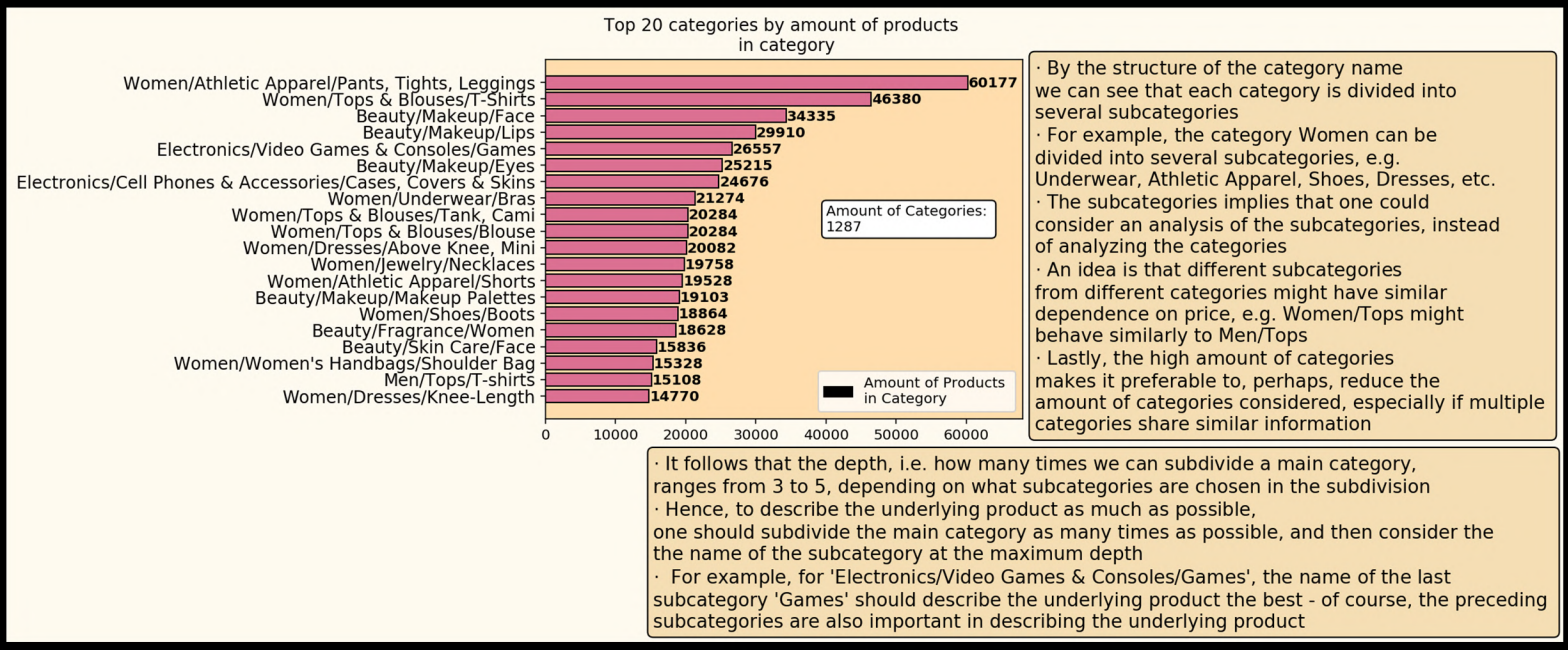


· 1: New
· 2: Almost New
· 3: Good
· 4: Fair
· 5: Poor

· 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 - Categories and the main categories of the dataset

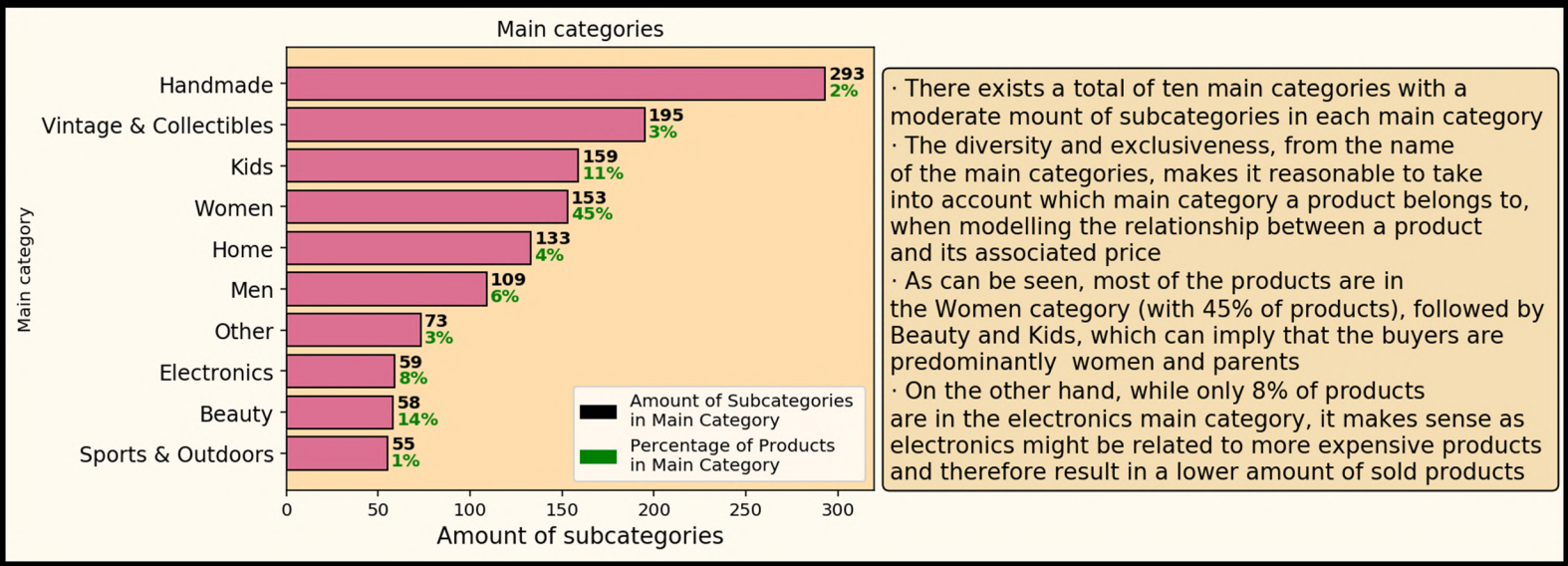
- For an idea of what type of categories each product belongs to, we consider an analyze of the category attribute -What categories exists? What are their names?



· By the structure of the category name we can see that each category is divided into several subcategories
· For example, the category Women can be divided into several subcategories, e.g. Underwear, Athletic Apparel, Shoes, Dresses, etc.
· The subcategories implies that one could consider an analysis of the subcategories, instead of analyzing the categories
· An idea is that different subcategories from different categories might have similar dependence on price, e.g. Women/Tops might behave similarly to Men/Tops
· Lastly, the high amount of categories makes it preferable to, perhaps, reduce the amount of categories considered, especially if multiple categories share similar information

· It follows that the depth, i.e. how many times we can subdivide a main category, ranges from 3 to 5, depending on what subcategories are chosen in the subdivision
· Hence, to describe the underlying product as much as possible, one should subdivide the main category as many times as possible, and then consider the the name of the subcategory at the maximum depth
· For example, for 'Electronics/Video Games & Consoles/Games', the name of the last subcategory 'Games' should describe the underlying product the best - of course, the preceding subcategories are also important in describing the underlying product

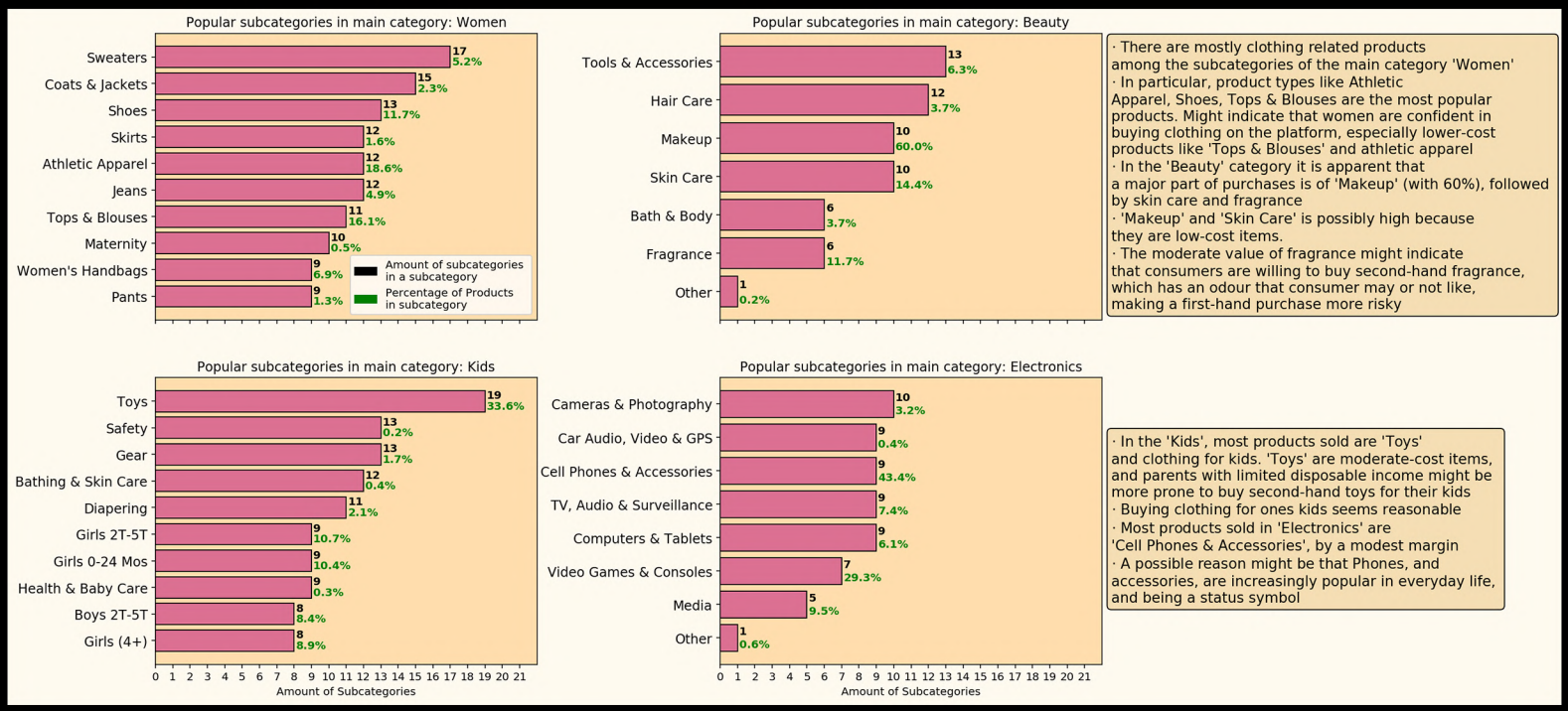
- To continue, we consider the different main categories that exists, to get a rough idea of what type of products exists



· There exists a total of ten main categories with a moderate mount of subcategories in each main category
· The diversity and exclusiveness, from the name of the main categories, makes it reasonable to take into account which main category a product belongs to, when modelling the relationship between a product and its associated price
· As can be seen, most of the products are in the Women category (with 45% of products), followed by Beauty and Kids, which can imply that the buyers are predominantly women and parents
· On the other hand, while only 8% of products are in the electronics main category, it makes sense as electronics might be related to more expensive products and therefore result in a lower amount of sold 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:

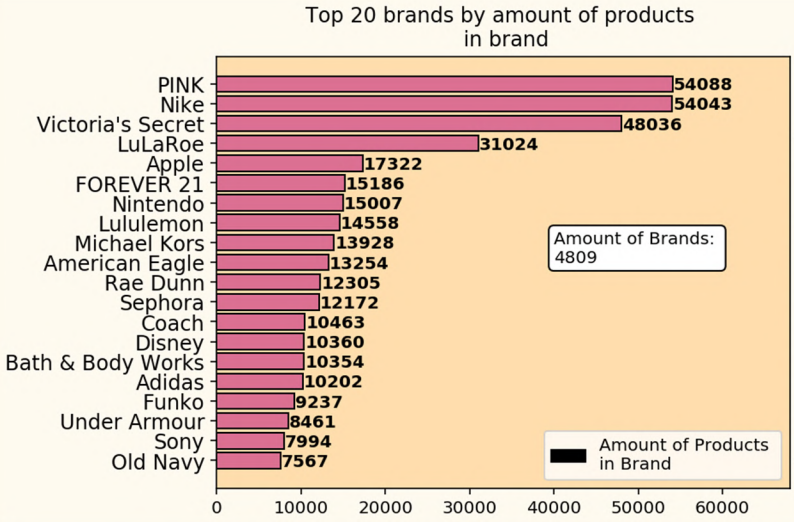
· Electronics/Computers & Tablets/iPad/Tablet/eBook Access
· Electronics/Computers & Tablets/iPad/Tablet/eBook Readers

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

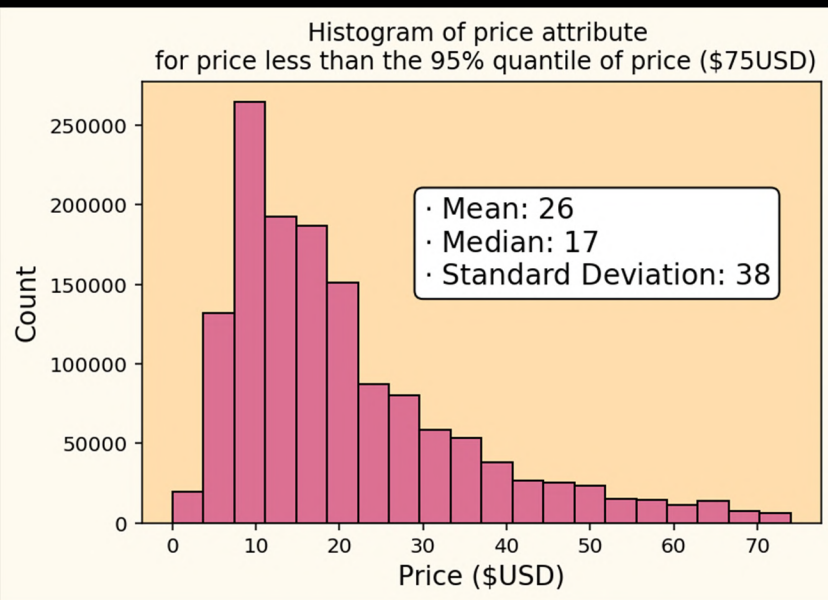
Mercari Price Analysis - Data Analysis - Brands of products, and the price of products

- 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



· A lot of popular brands are present, like Nike, Apple, Michael Kors, Nintendo, Disney
· A total of 4809 brands indicates that it might be a good idea to analyze the average price of products in brands - to quantitatively compare the price-class of different brands, and perhaps relate certain brands to each other
· The ubiquity of popular brands is not a surprise, as there should naturally be a high demand for such products
· Another observation is how Nike is much more popular than Adidas, which is a competitor. Indicating that Nike products is to go-to sports brand for most consumers

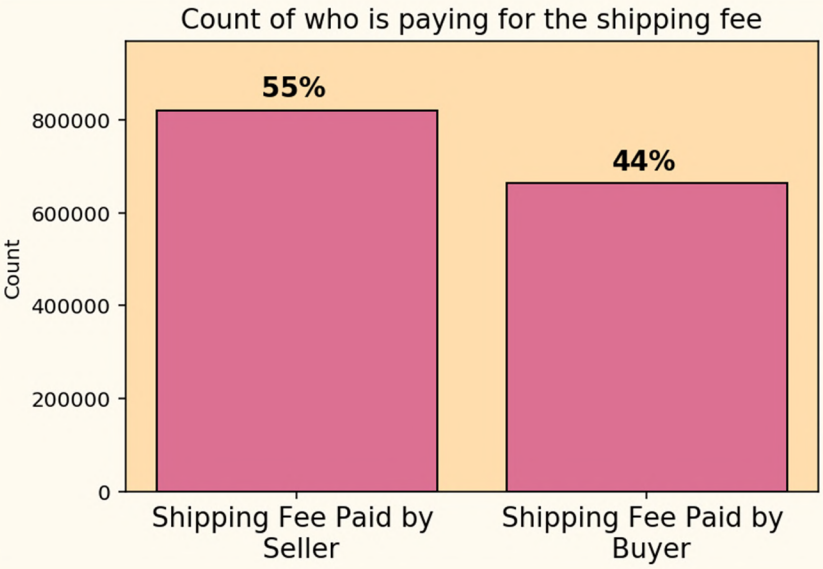
- The attribute of most interest is the price attribute, which is the price of the underlying product.



· A major part of products are bought at around 5-25 $USD - indicates that people predominantly buy low-cost products
· For increasing prices, the amount of products bought decreases almost exponentially - while high-cost items are bought, they are bought significantly less frequent compared to low-cost products
The 5-25$USD region could correspond to clothes, accessories, and other low-cost products
· It is possible that the high frequency of low-cost items bought might mean that consumers are less confident in buying high-cost products - since high-cost products might be more risky to buy second-hand rather than first-hand

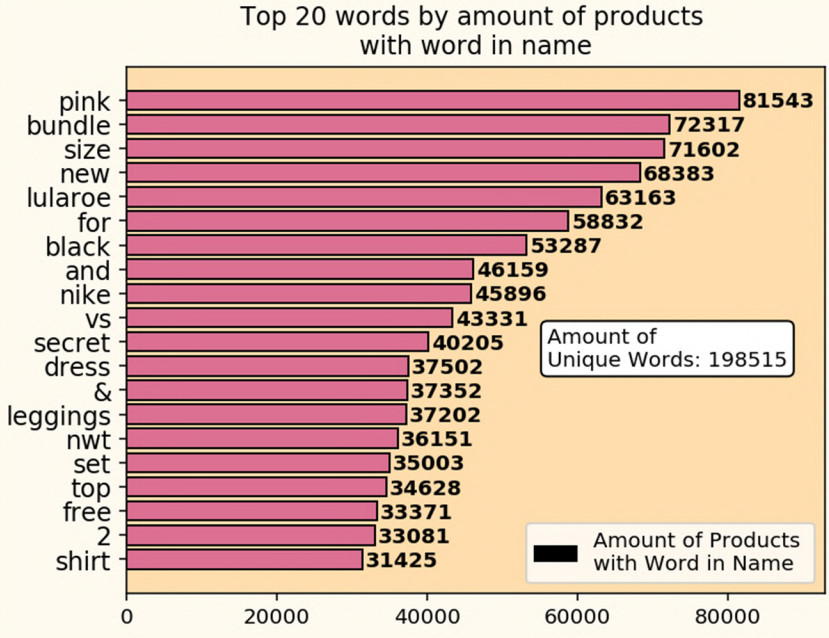
Mercari Price Analysis - Data Analysis - Shipping fee and the name of each product

- 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?



· 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

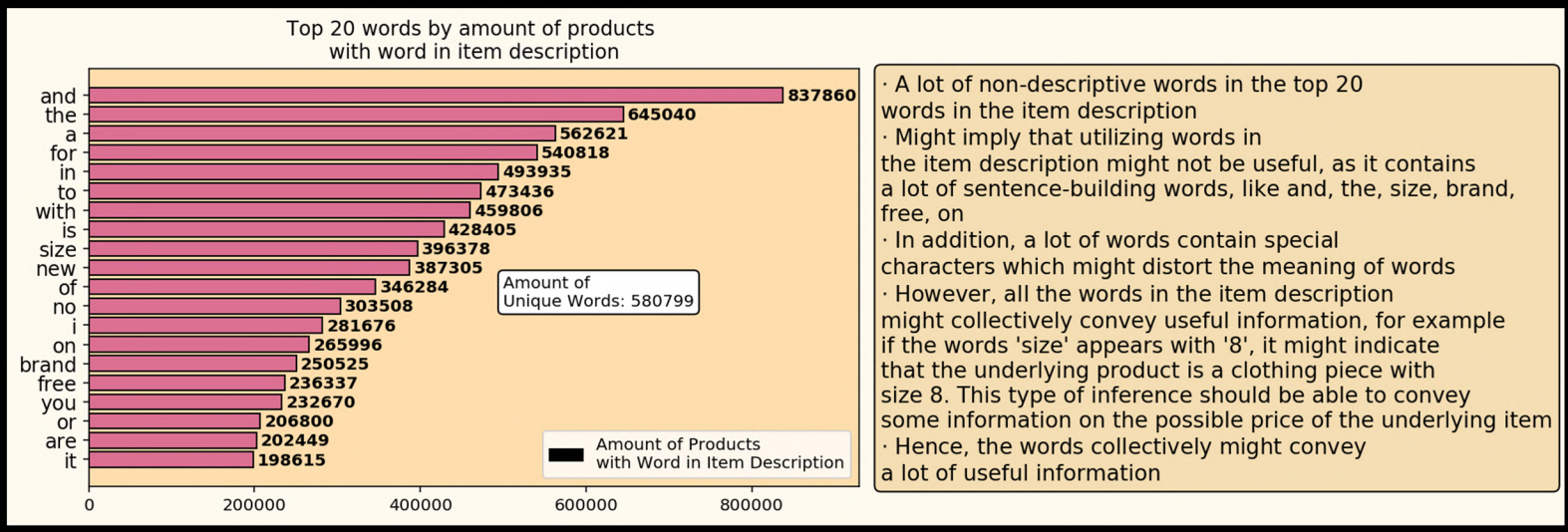
- Each product has an associated name, which is used to present the product listing to the potential buyer. Potentially, the words used in the name can be of interest in modelling the product's underlying price



· 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

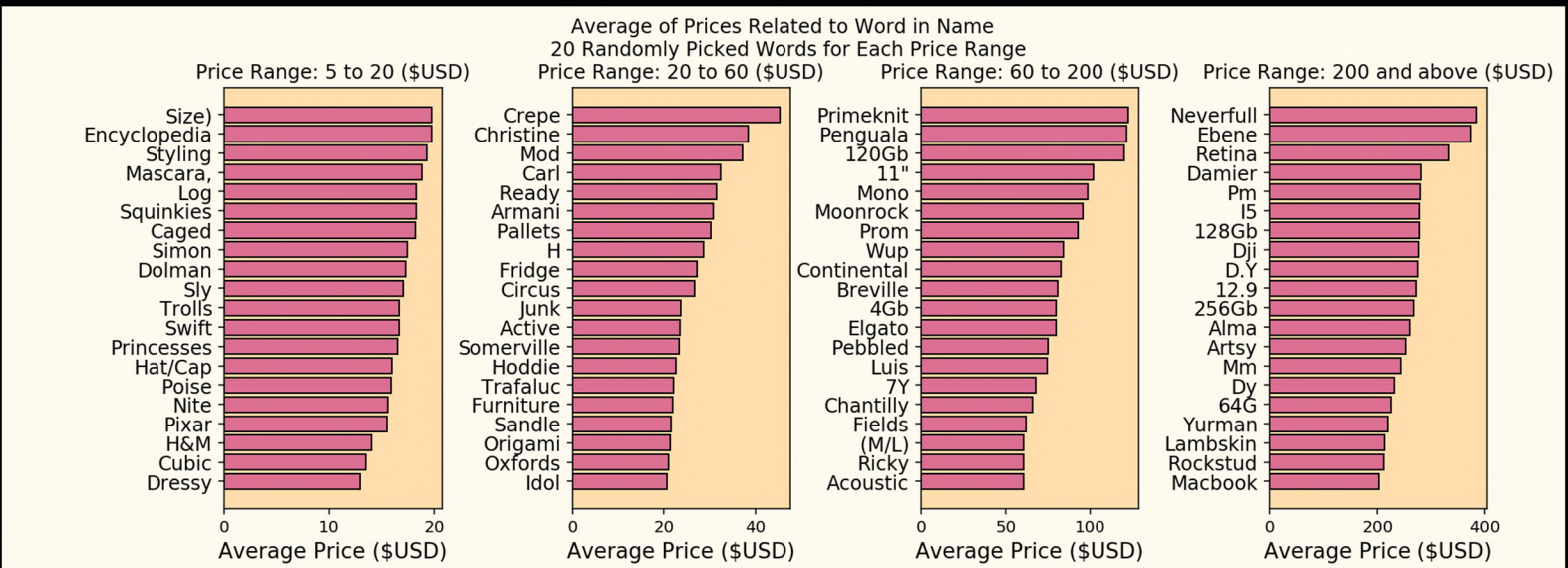
Mercari Price - Data Analysis - Words in item description and the average price associated with words in the name

- 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



· 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

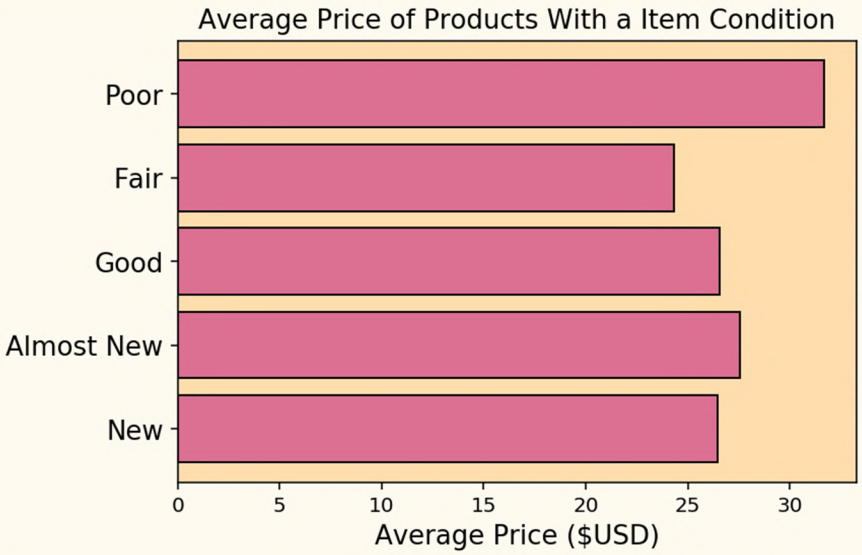


· A lot of descriptive words in all price ranges, including words like Trolls, Princesses, Armani, Hoddie, Oxfords, Acoustic, 120Gb, 256Gb, Macbook, Lambskin, Damier
· Some words give an almost full description of the underlying product, e.g. Macbook, while some words give an important aspect, e.g. Lambskin
· 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

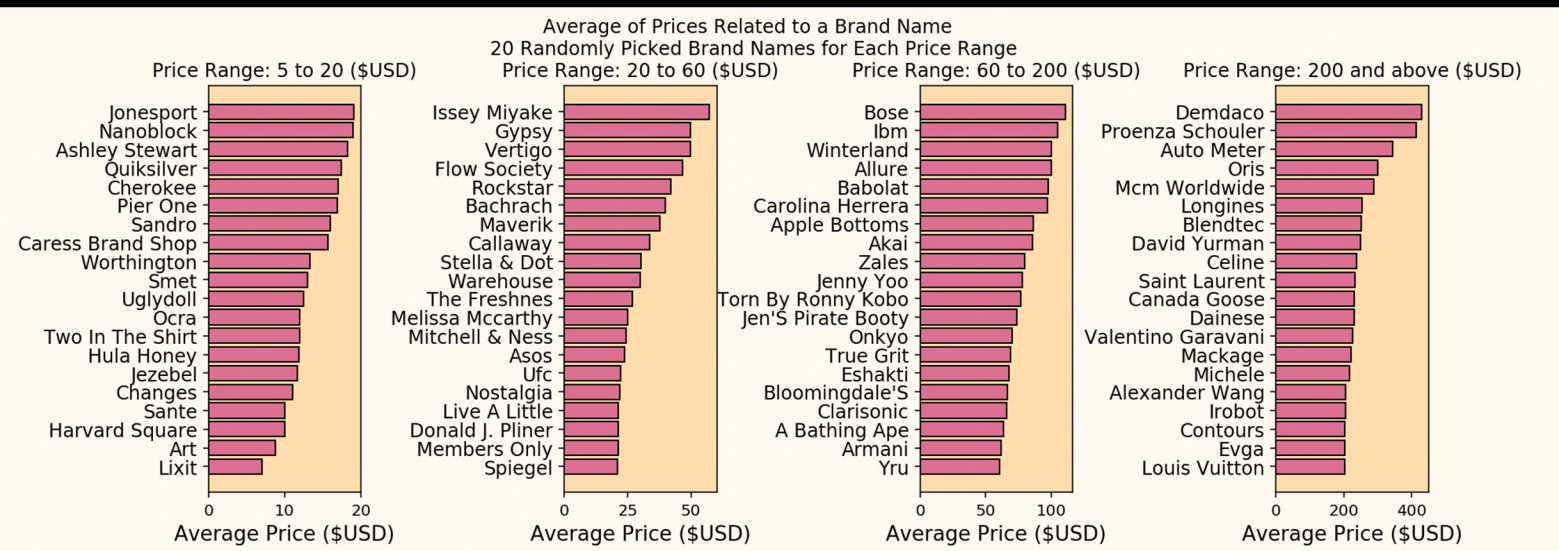
- A minor interest is to consider the prices of products with different item conditions - are less quality items generally cheaper?



· No particular pattern in the average price for different item conditions, as they are all pretty close to each other
· This suggest that it may be preferable to consider the item condition as a category when modelling the relationship between price and products

· For all possible Item Condition:
  Collect all products with that Item Condition
  Calculate the average price of all those products

- 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

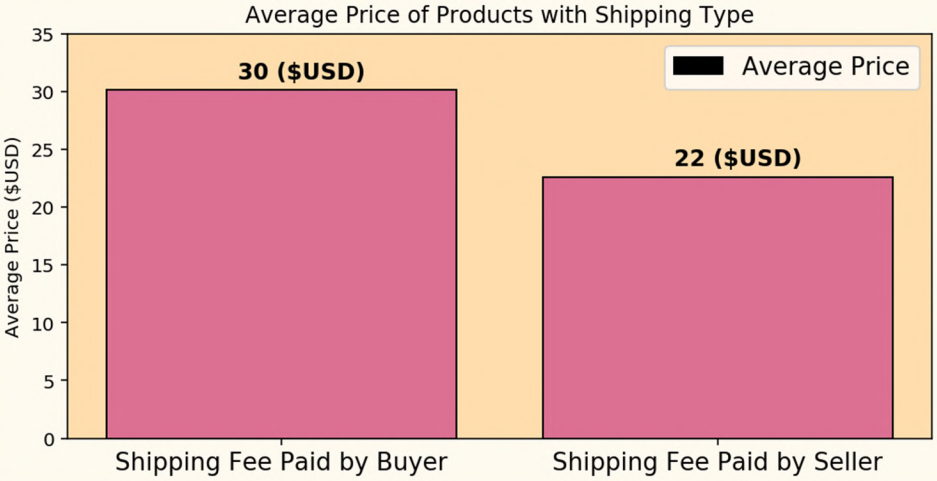


· A lot of known brands in all price ranges, including brands like Asos, David Yurman, Louis Vuitton, Saint Laurent, Alexander Wang
· The average price of products associated with a brand can give a rough idea of the price-class associated with the brand - some may be expensive like Louis Vuitton while others may be more affordable like Pier One
· In addition, certain brands will have a similar price-class, and will probably sell products for similar prices

· Construction of the Average Price:
For each possible Brand Name:
  Collect all products associated with that Brand Name
  Calculate the average price of all those products

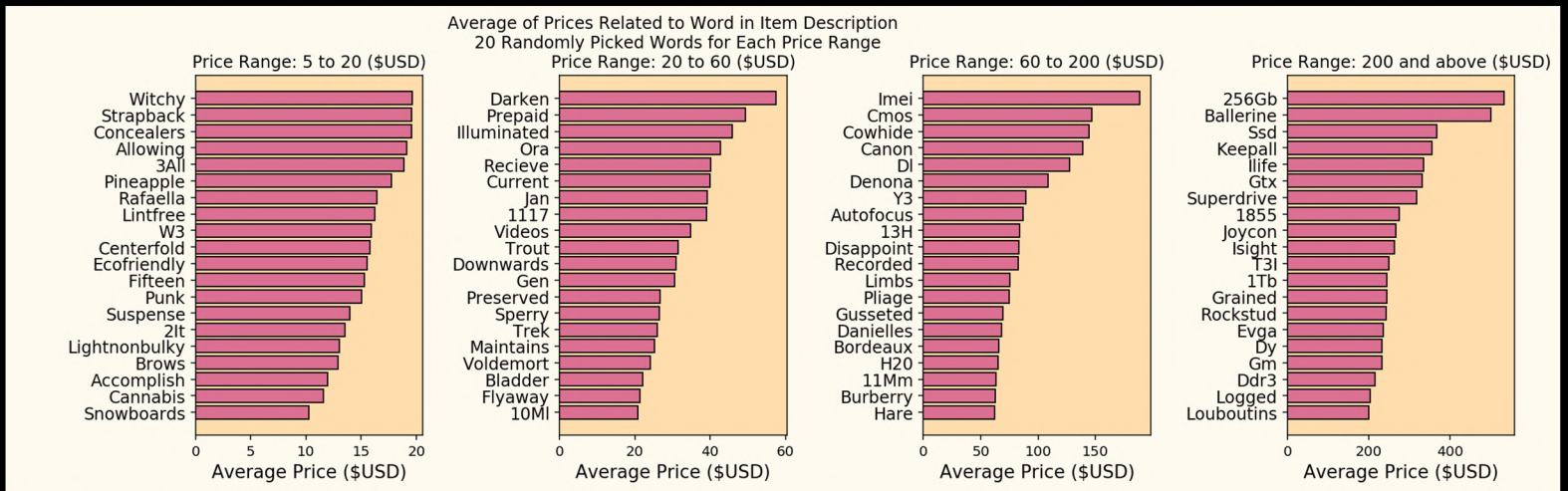
Mercari Price Analysis - Data Analysis - Shipping fee with price, and the price associated with words in the item description

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



· There is a small difference in the average price between the two different shipping types
· The difference is relatively small, but indicates that products for which the shipping is paid by the buyer is a little bit more expensive
· It should be noted that the Price attribute does not include the shipping fee price paid by, when that is the case, a potential buyer

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

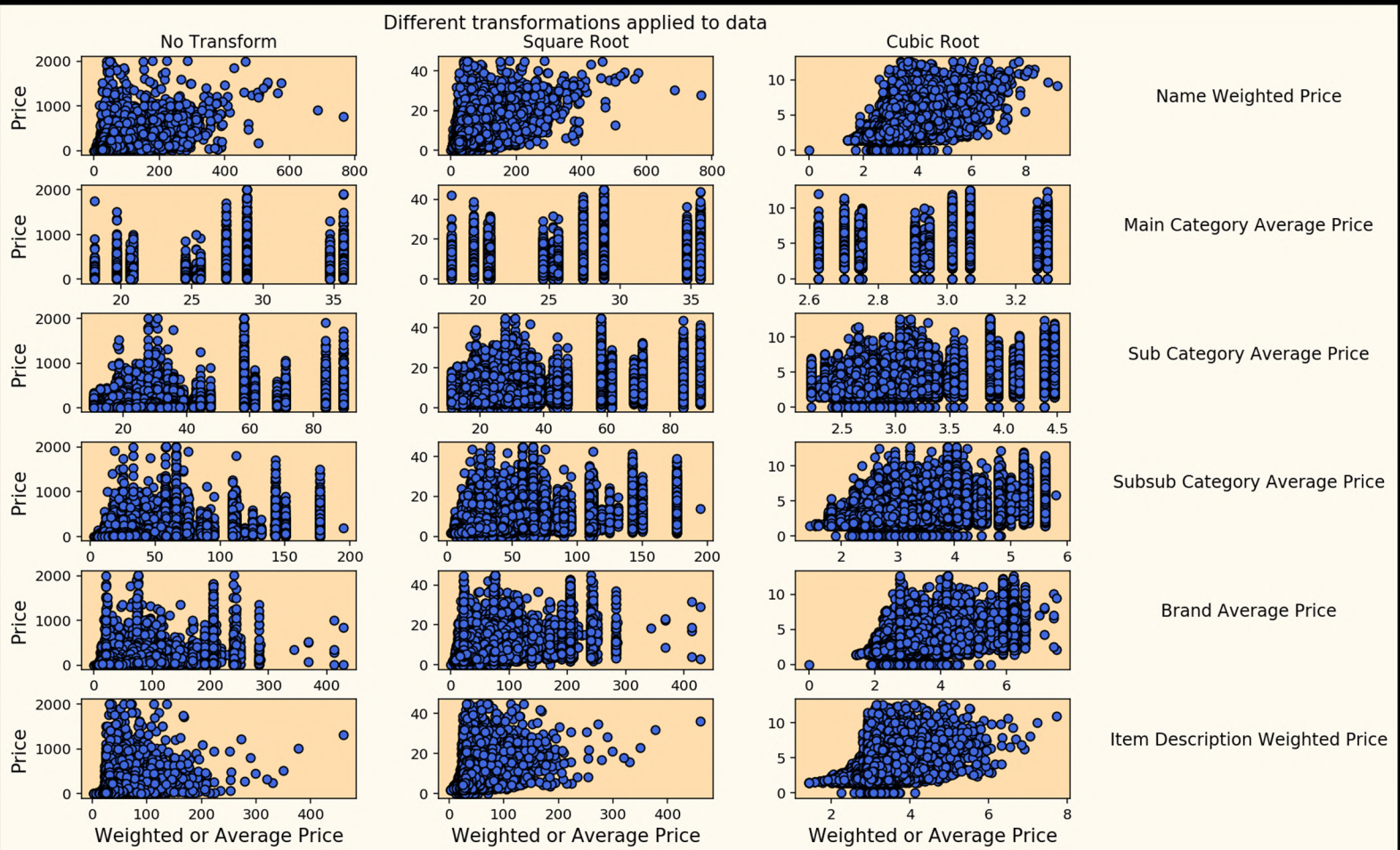


· The words in the item descriptions range from being descriptive, like Illuminated, Ecofriendly, Grained, Voldemort, to less descriptive, like Disappoint, Gen, Recieve. In general, the words seems to indicate some characteristic of the underlying product. For example, a combination of Ecofriendly and Burberry might indicate that the underlying product is not only of the luxury brand Burberry, but also an ecofriendly product· For a given product, all the words in its item description should collectively, with the average prices defined above, give a rough indication of a reasonable, possible price of the underlying product, for example by weighing the average prices of each word
· A positive thing about the item description is that it may convey more indepth information of the underlying product as compared to the name attribute of the product

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

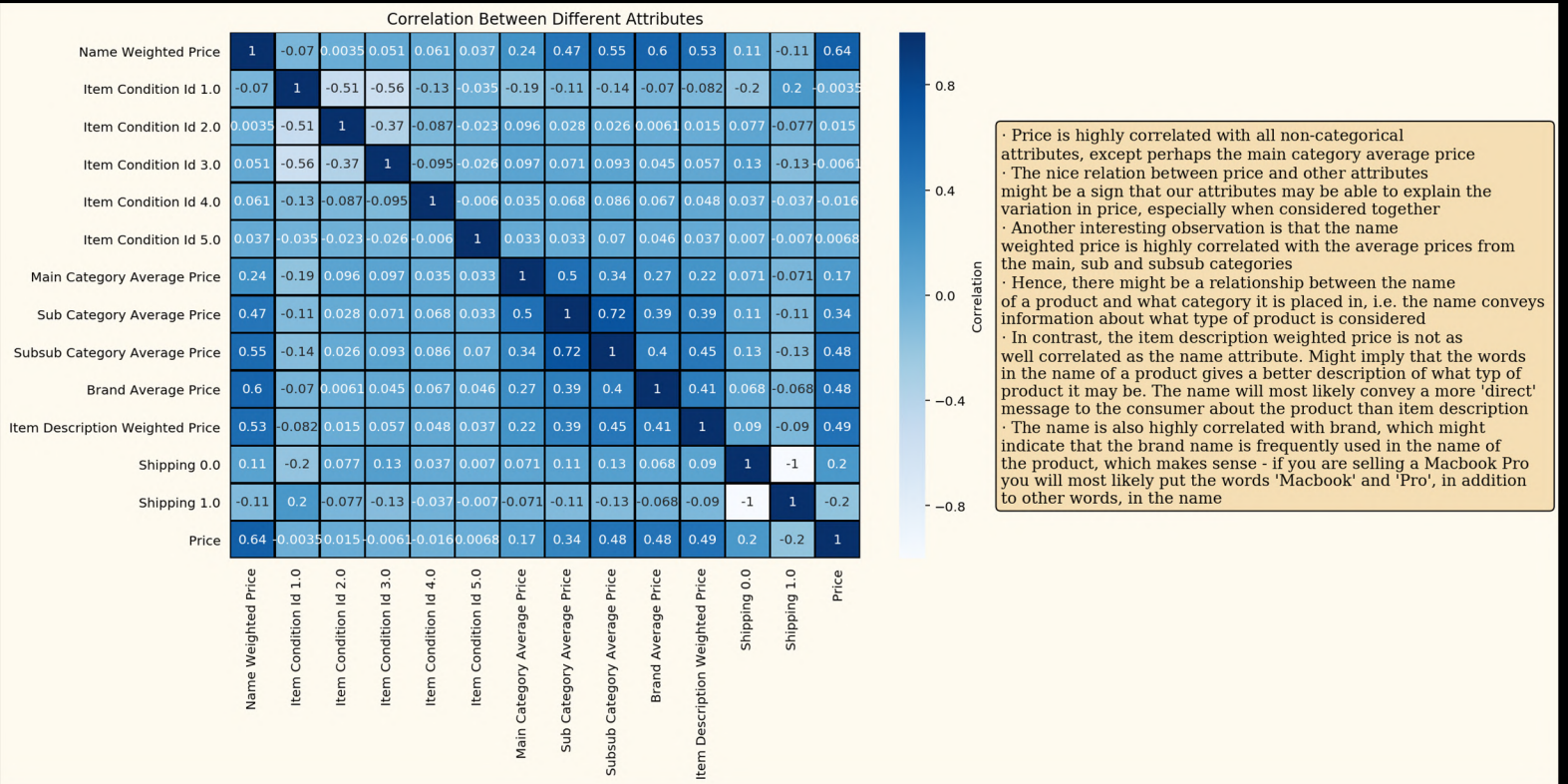
Mercari Price Analysis - Prediction Analysis - Transformation of variables, correlation among variables

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



· The cubic root transformation makes the different attributes a bit more linearly related with the price attribute
· On the other hand, one can see that the large amount of data points (above 1 million) makes the relationships really crowded

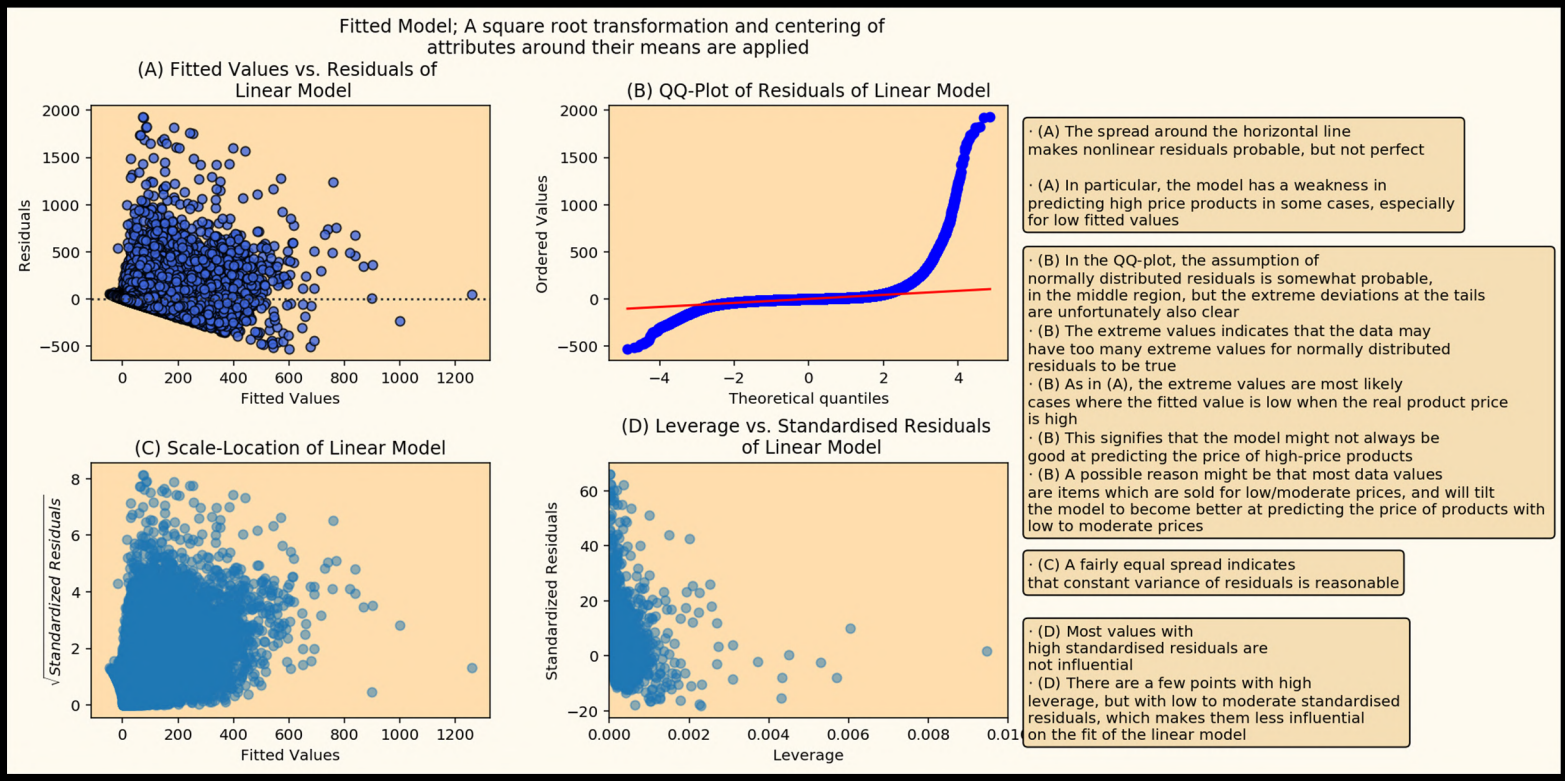
- A key idea in linear models is to analyze the correlation among all variables considered, especially the correlations with respect to the price variable



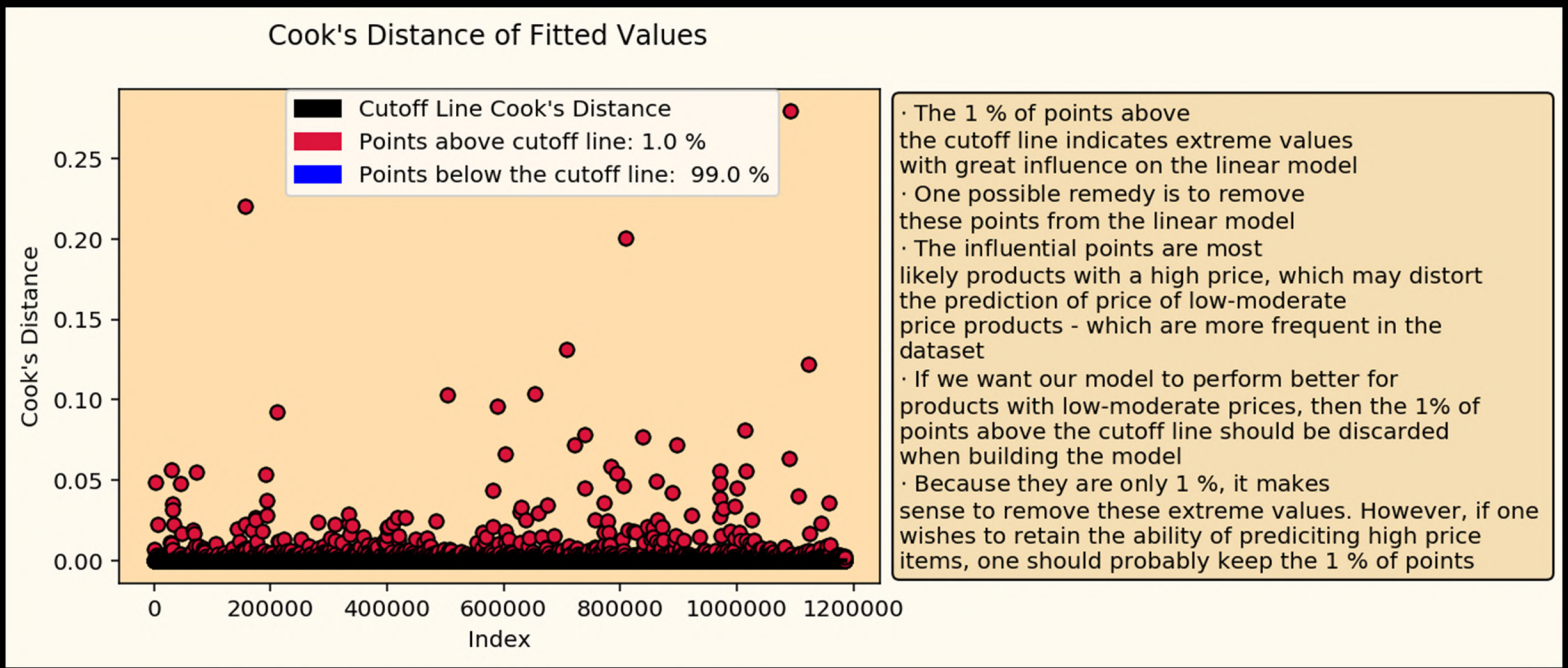
· Price is highly correlated with all non-categorical attributes, except perhaps the main category average price
· The nice relation between price and other attributes might be a sign that our attributes may be able to explain the variation in price, especially when considered together
· Another interesting observation is that the name weighted price is highly correlated with the average prices from the main, sub and subsub categories
· Hence, there might be a relationship between the name of a product and what category it is placed in, i.e. the name conveys information about what type of product is considered
· In contrast, the item description weighted price is not as well correlated as the name attribute. Might imply that the words in the name of a product gives a better description of what typ of product it may be. The name will most likely convey a more 'direct' message to the consumer about the product than item description
· The name is also highly correlated with brand, which might indicate that the brand name is frequently used in the name of the product, which makes sense - if you are selling a Macbook Pro you will most likely put the words 'Macbook' and 'Pro', in addition to other words, in the name

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

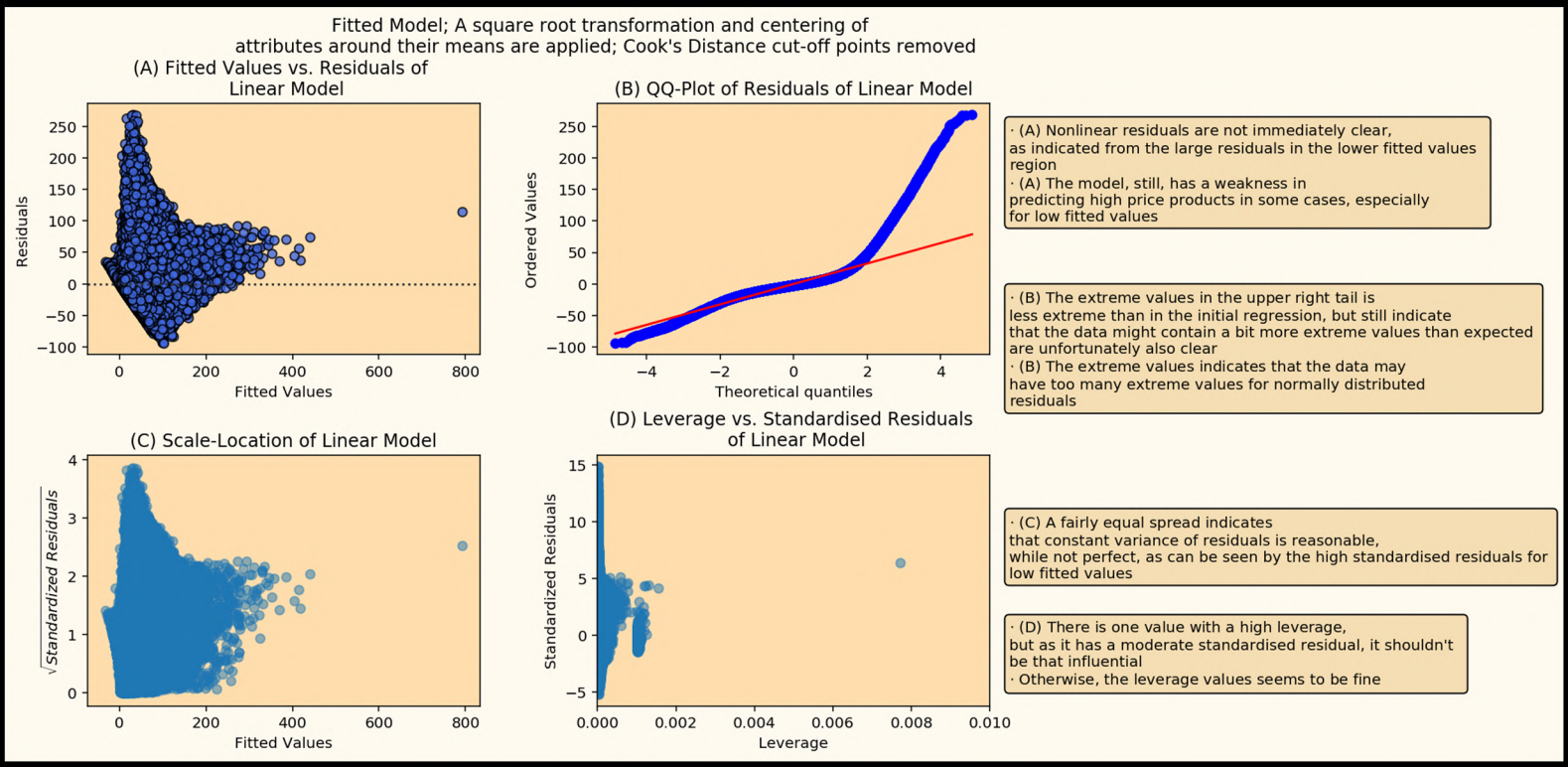


- Further describing the linear model, especially illustrating data points with influence on the fit of the model, the Cook's Distance can be considered

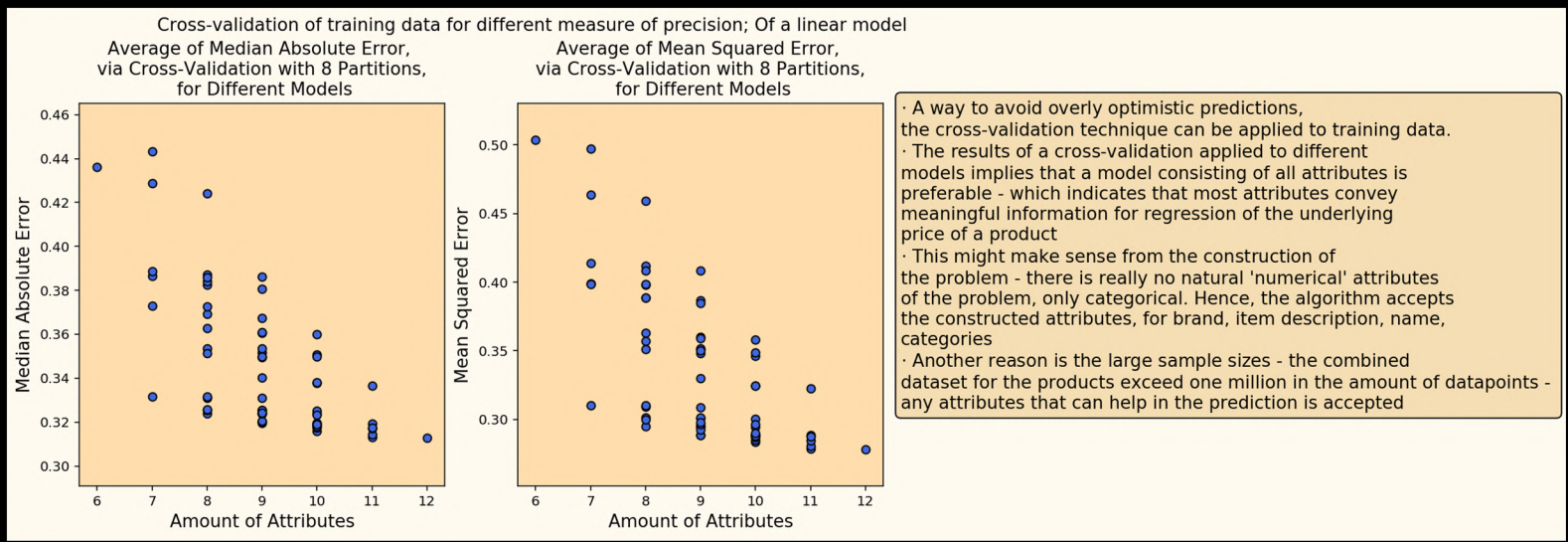


Mercari Price Analysis - Prediction Analysis - Linear model after Cook's Distance analysis, and different performance measures to choose an optimal model

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

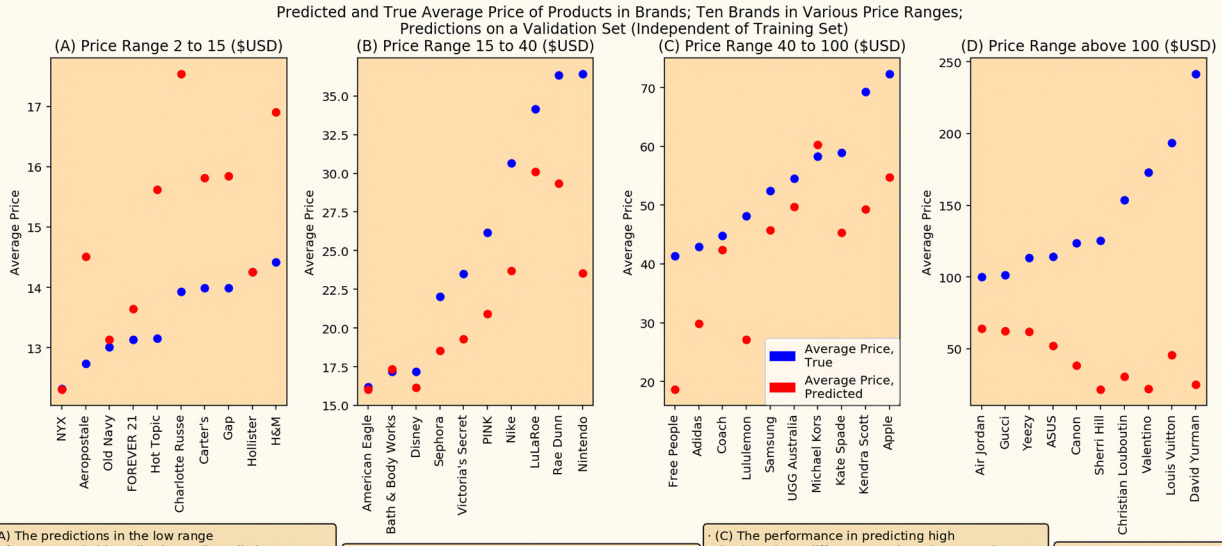


- The aim is to find an optimal model. For this, we consider all combinations of attributes (where all binary attributes are kept in the model), and for each combination we evaluate two performance measures. Lastly, based on the performances, we pick a final, optimal model



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



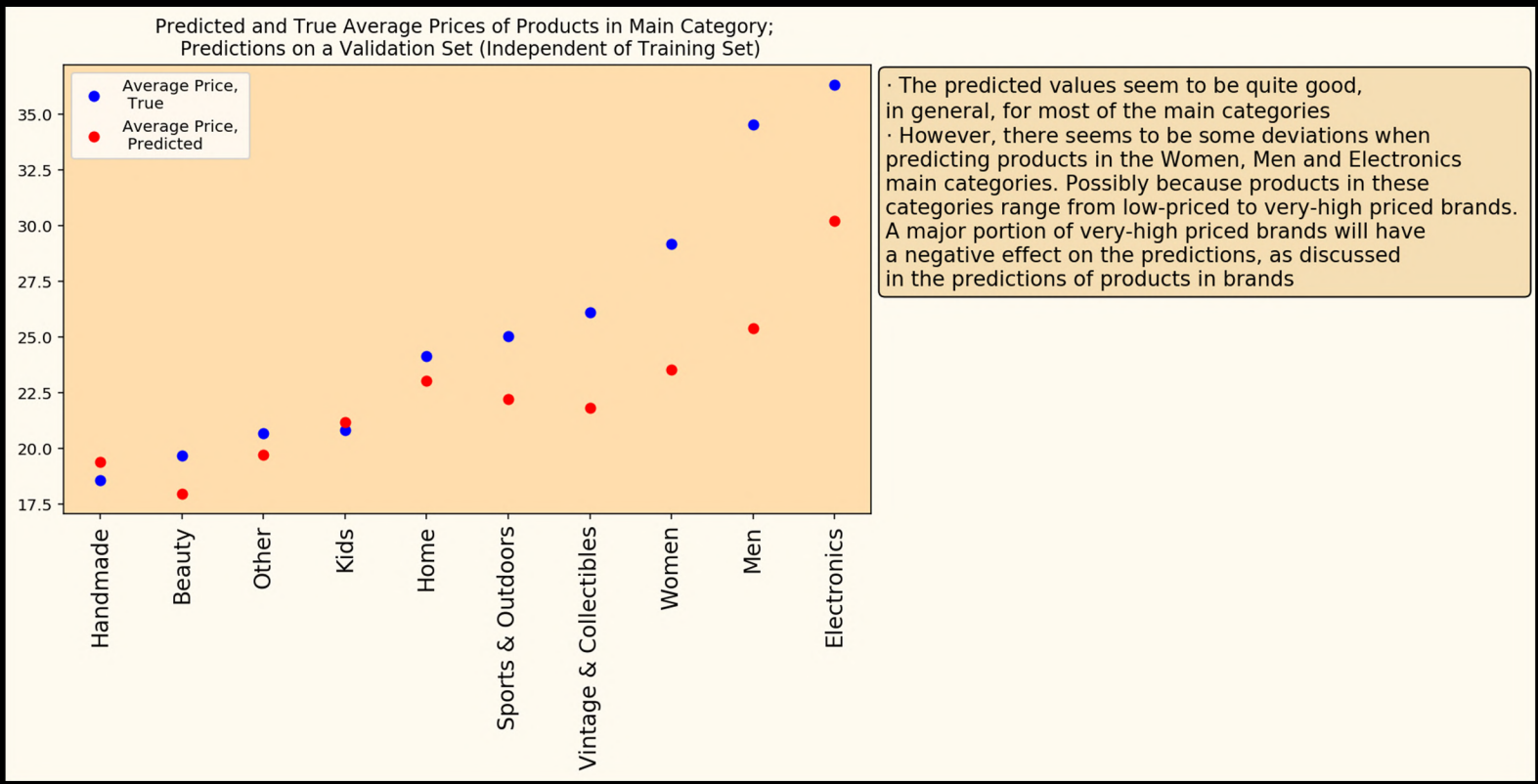
· (A) The predictions in the low range performs remarkably well, where all predictions are just a few dollars away from the true values.
· (A) Indicates that our model is in general good for predicting product prices for low-price range brands like H&M, Old Navy and Gap
· (A) A possible reason might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



· 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 correponds to a diverse set of products, and this will show how the model performs on a diverse set of categories