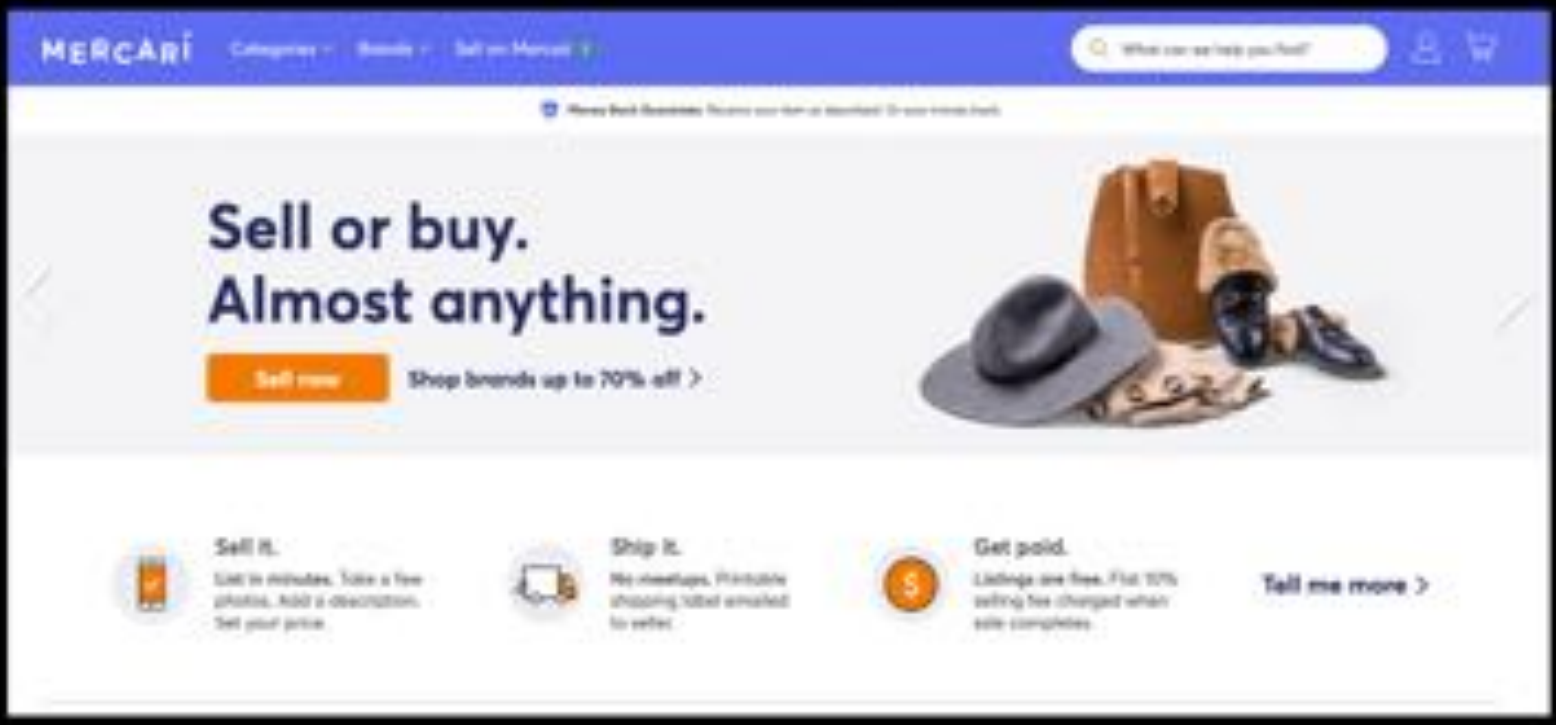
Mercari Price Analysis Project



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

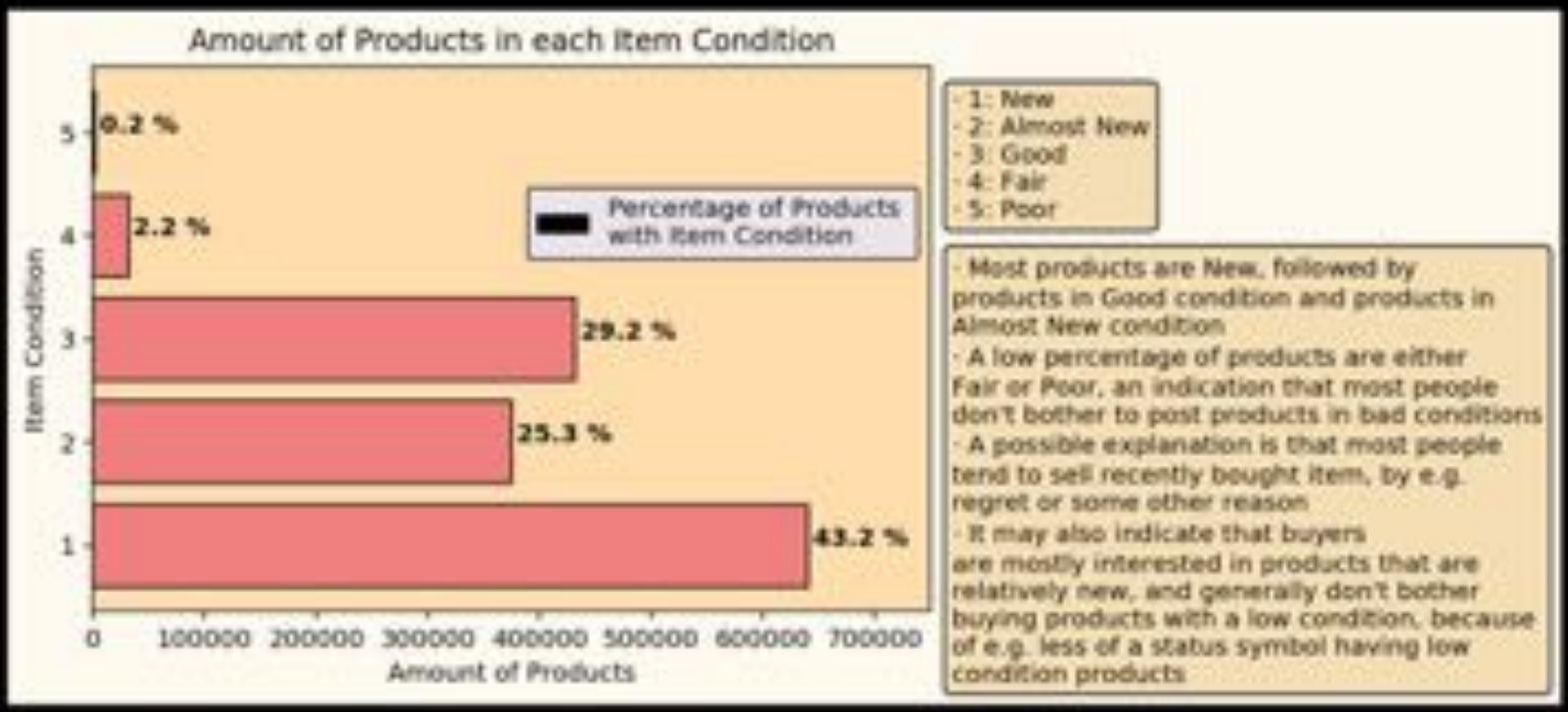
Mercari Price Analysis - Data Analysis - Description of dataset, item condition of products

- 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

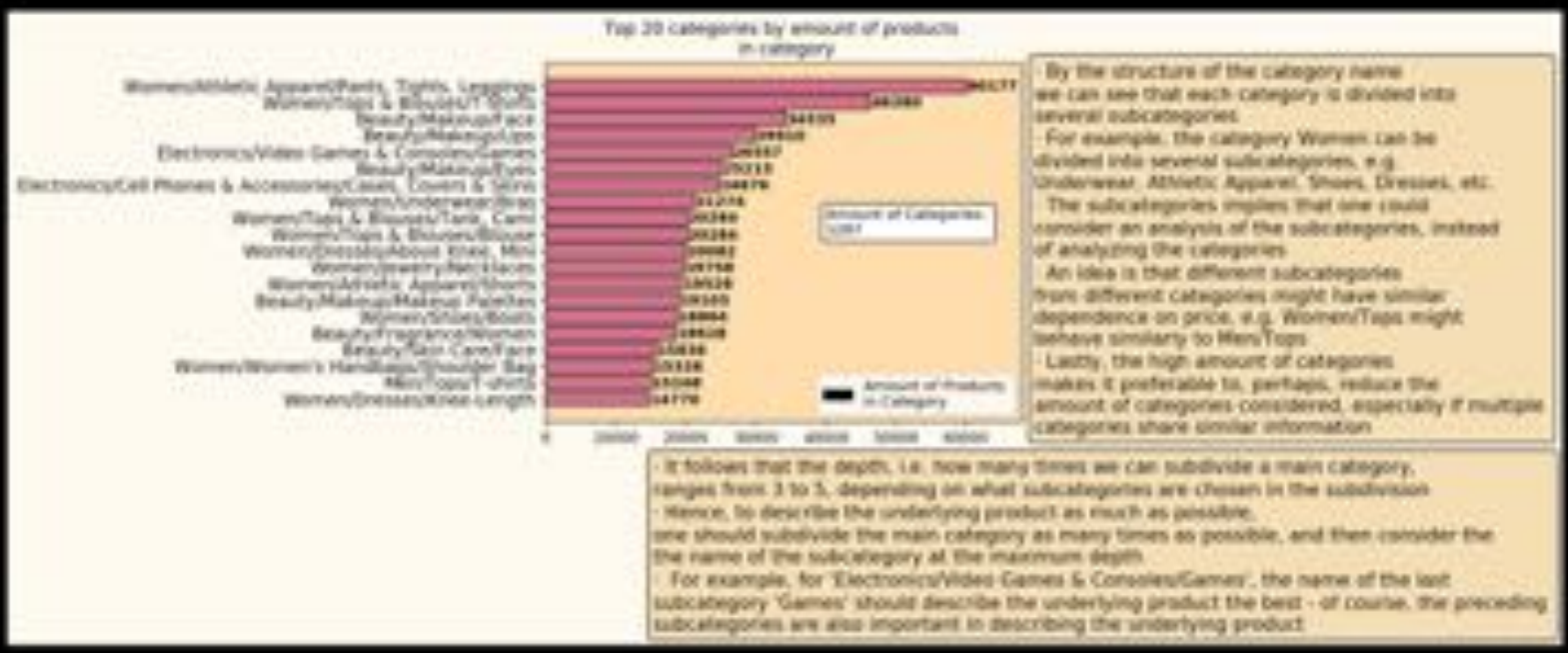
· The Name is a typical, brief description of the the product in question
· The Item Condition is a number representing the condition of the product in question
· The Category Name represents the category of the product
· The Brand Name is simply the brand of the underlying product, e.g. Nike
· The Price 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

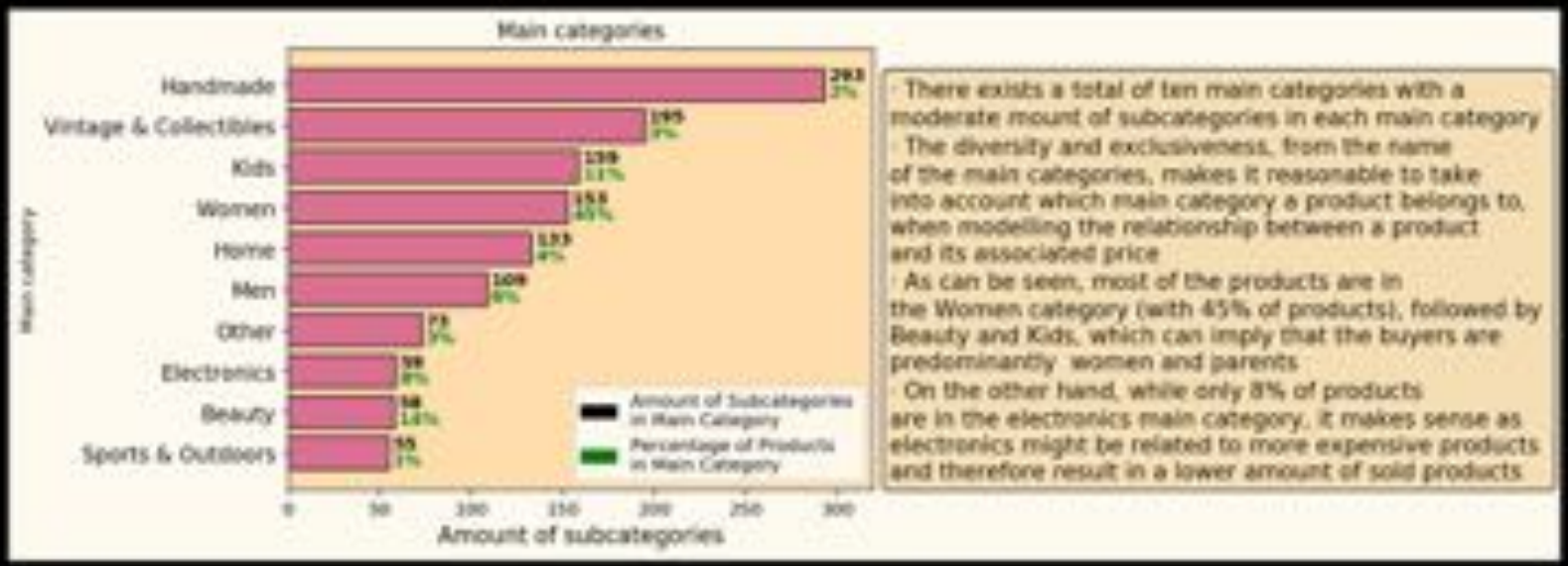


Mercari Price Analysis - Data Analysis - Categories and the main categories of the dataset

- 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 - 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:

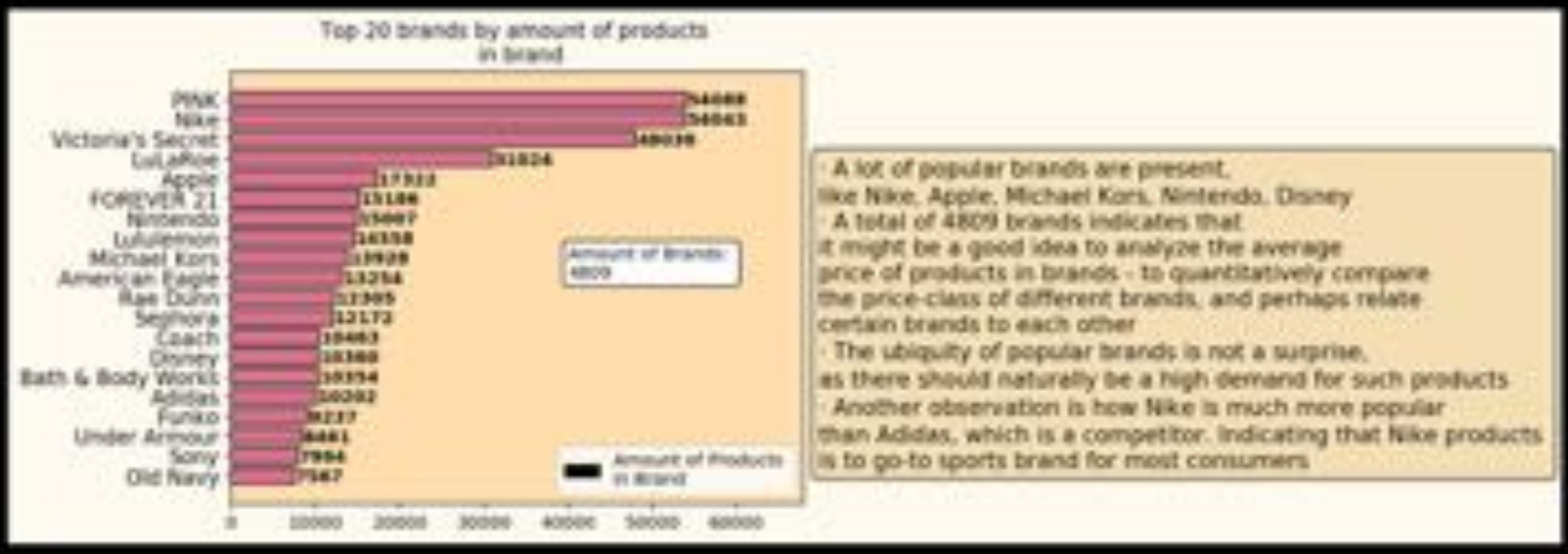
- Electronics/Computers & Tablets/iPad/Tablet/eBook Access
- Electronics/Computers & Tablets/iPad/Tablet/eBook Readers

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

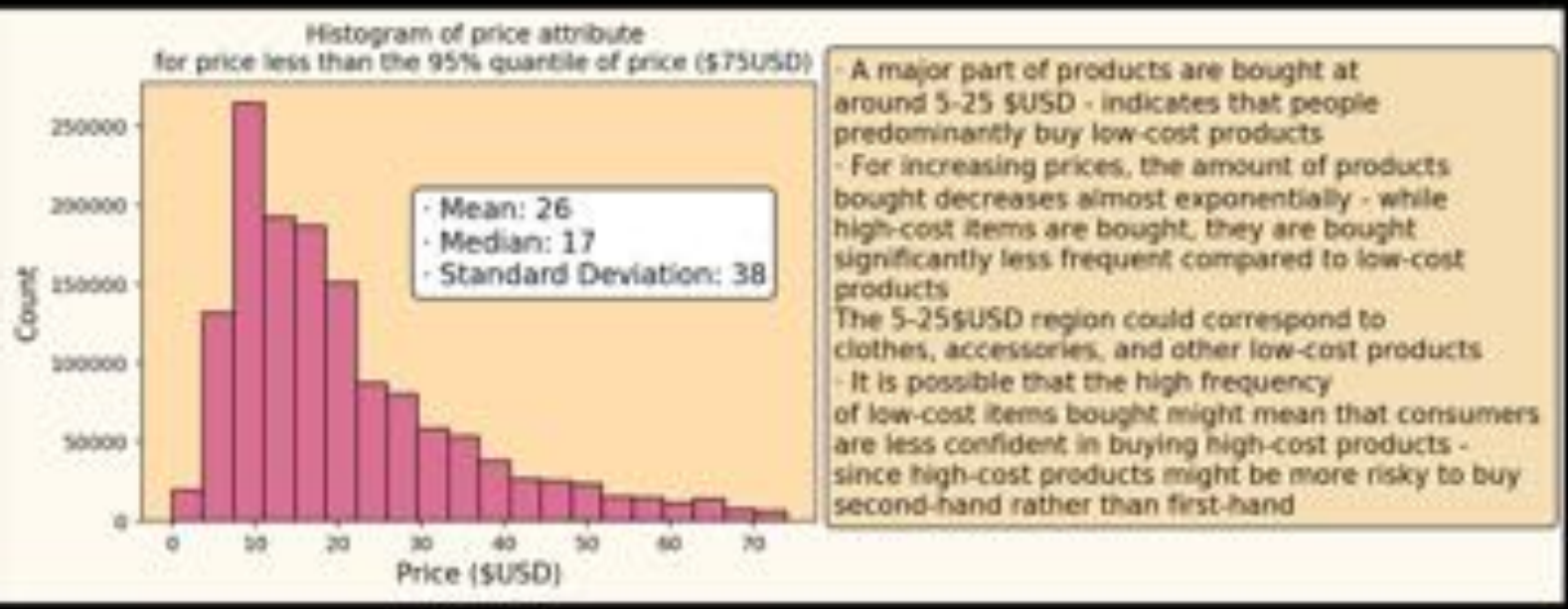
- Handmade/Housewares/Entertaining Serving
- Men/Coats & Jackets/Flight Bomber
- Men/Coats & Jackets /Varsity Baseball
- Sports & Outdoors/Exercise/Dance Ballet
- Sports & Outdoors/Outdoors/Indoor Outdoor Games
- Electronics/Computers & Tablets/iPad Tablet eBook Access
- Electronics/Computers & Tablets/iPad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



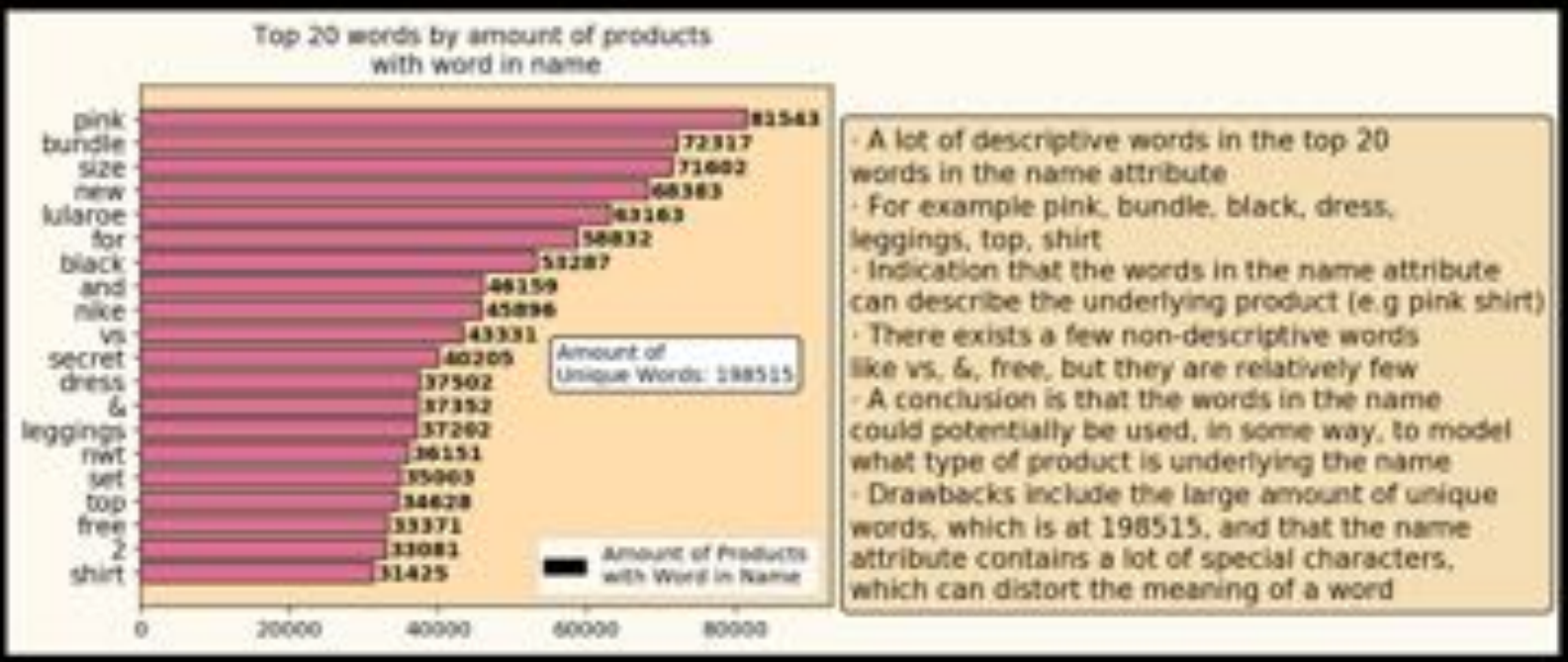
- The attribute of most interest is the price attribute, which is the price of the underlying product.



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

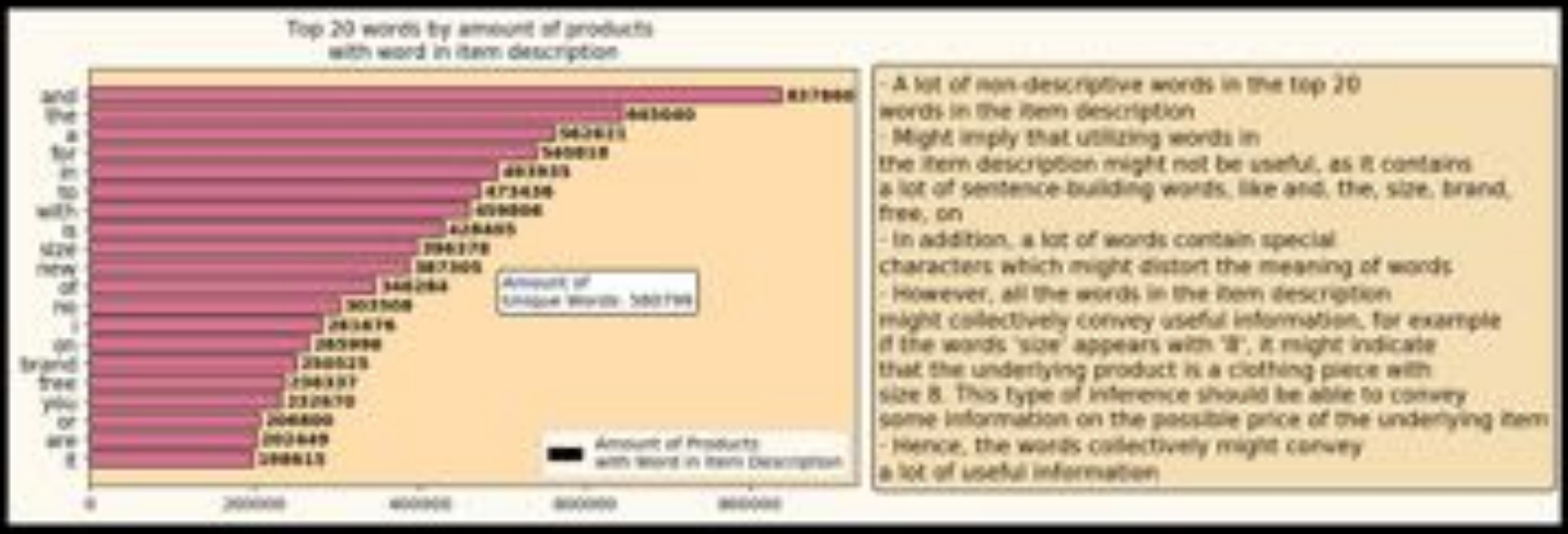
- For each product bought and sold on the platform, there is an associated shipping fee that is either paid by the buyer or the seller - what is its distribution?

- 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

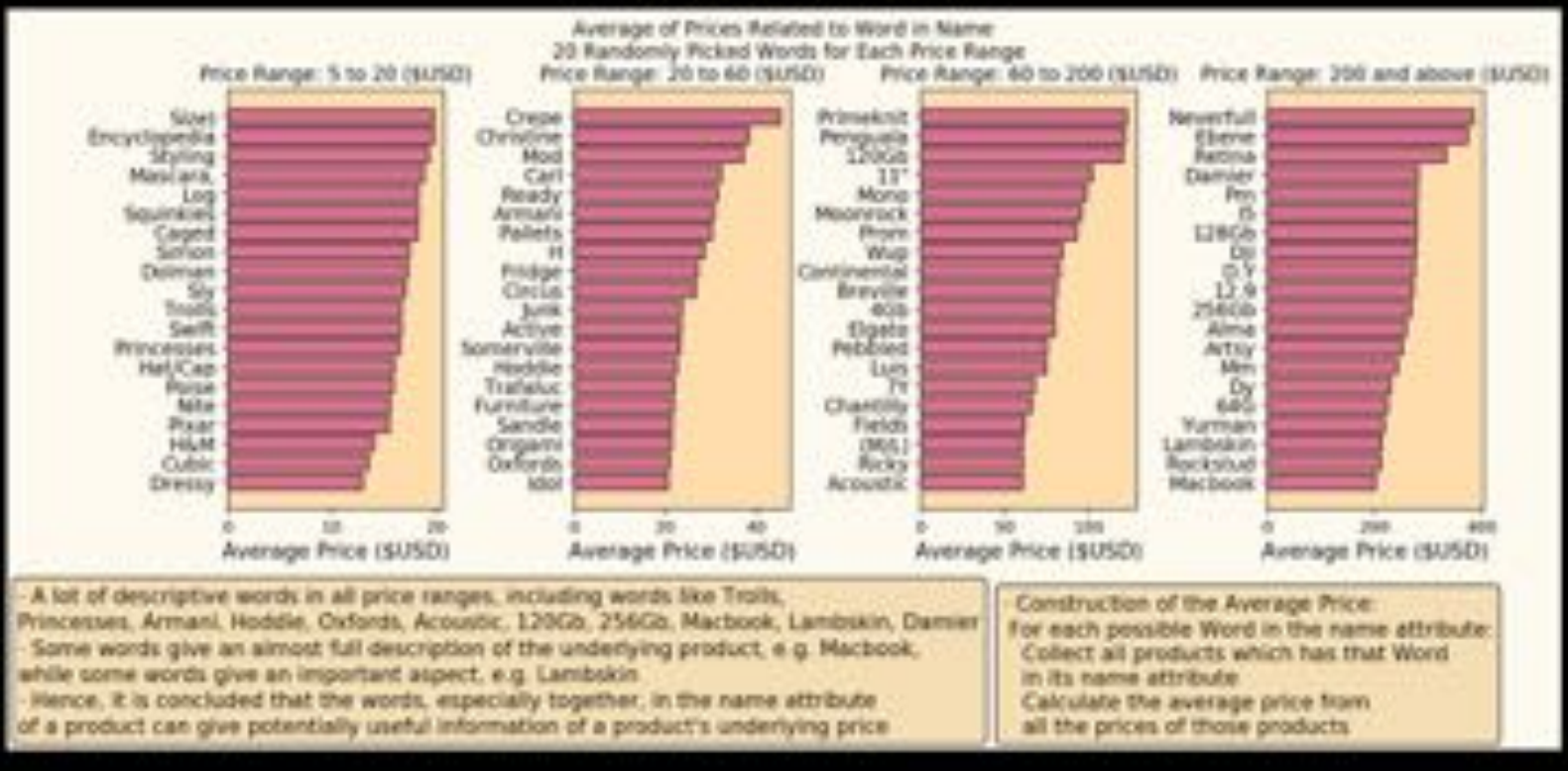


Mercari Price - Data Analysis - Words in item description and the average price associated with words in the name

- 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

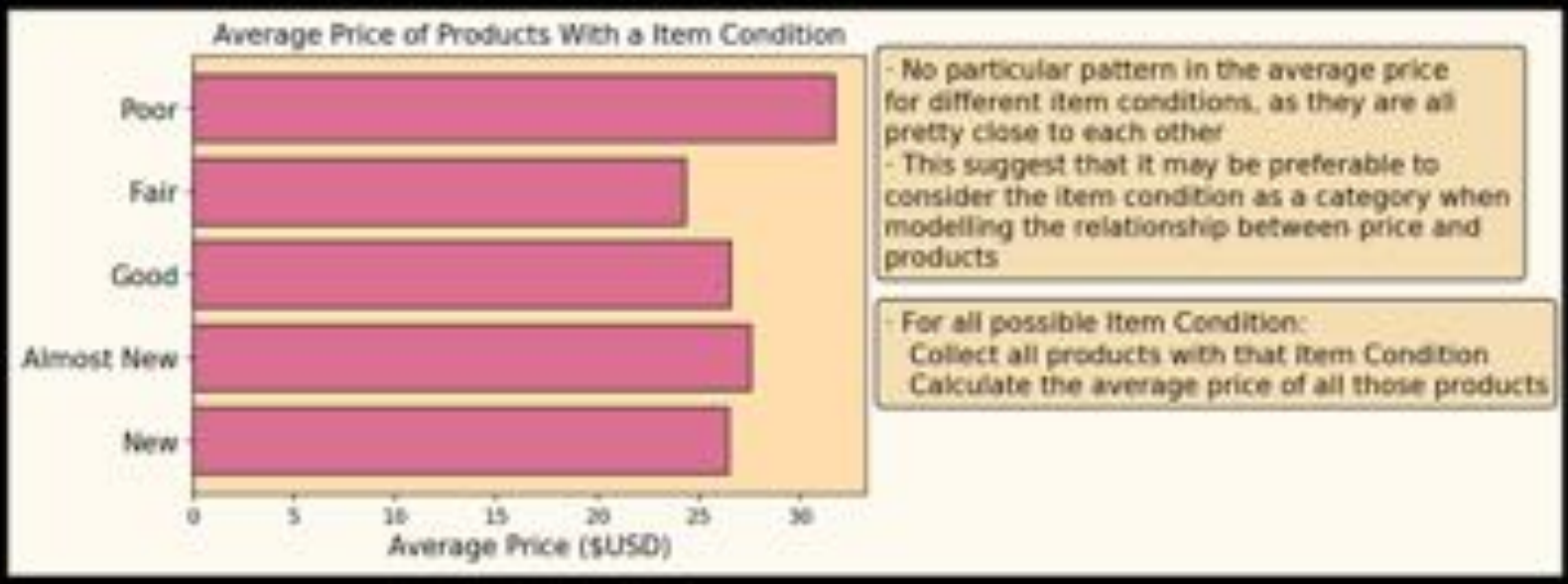


- In addition to what words exists in the name of products, there is an interest in correlating those words with prices of underlying products

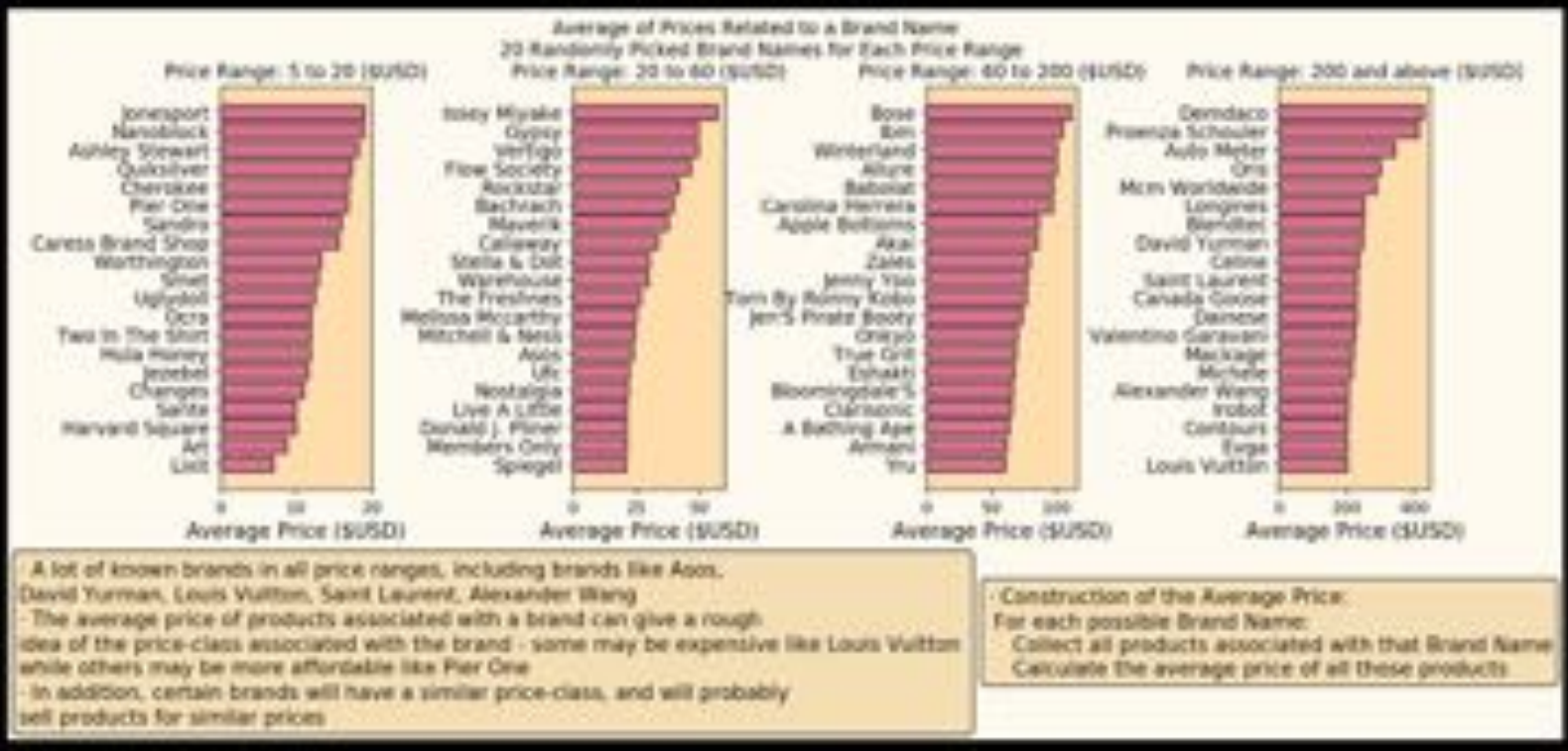


Mercari Price Analysis - Data Analysis - Prices associated with item condition of products, and prices of products in different brands

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



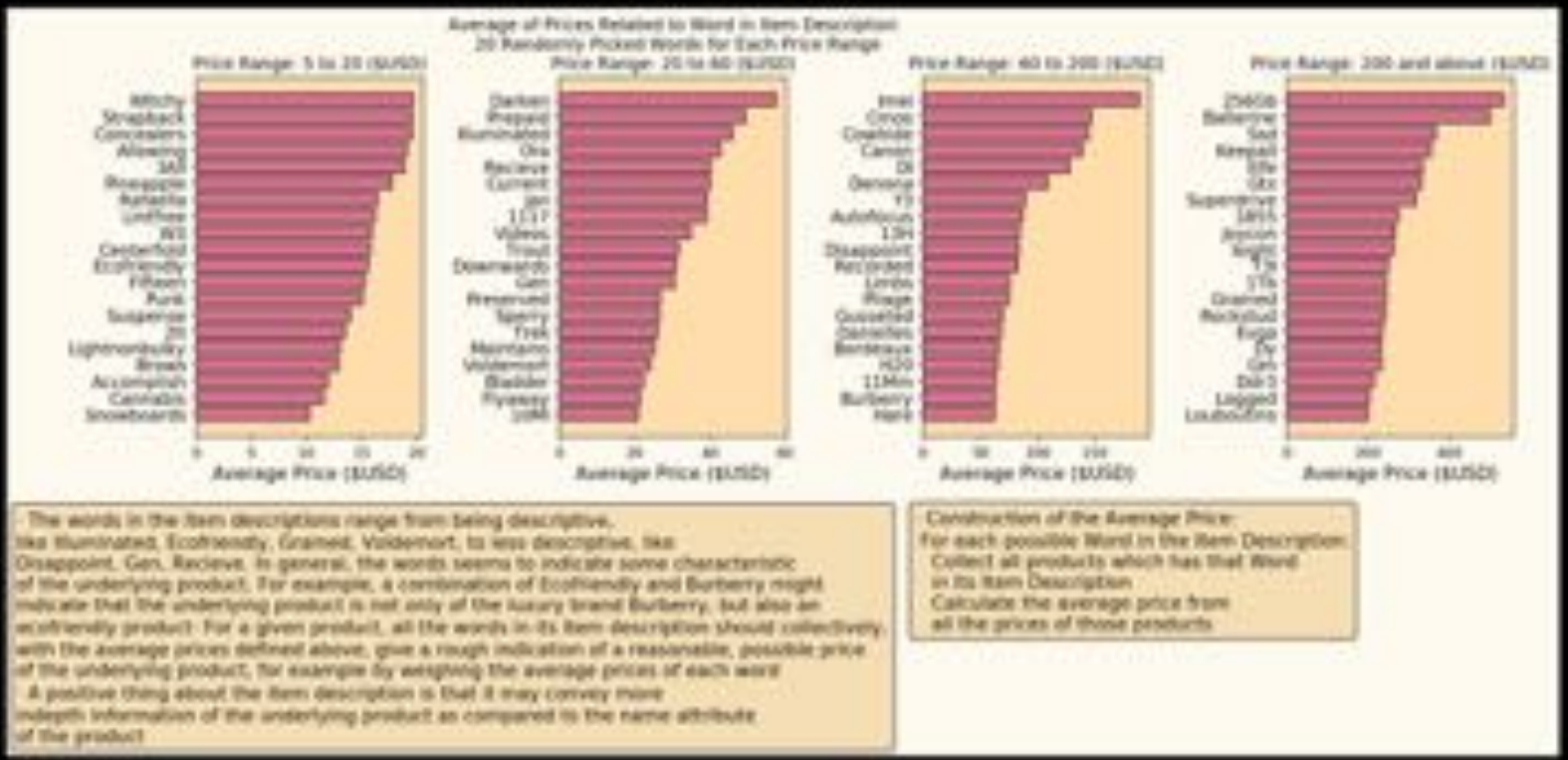
- A major factor of the price will obviously be which brand the underlying product belongs to - especially what type of price products in that brand usually sell for, in a general sense



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

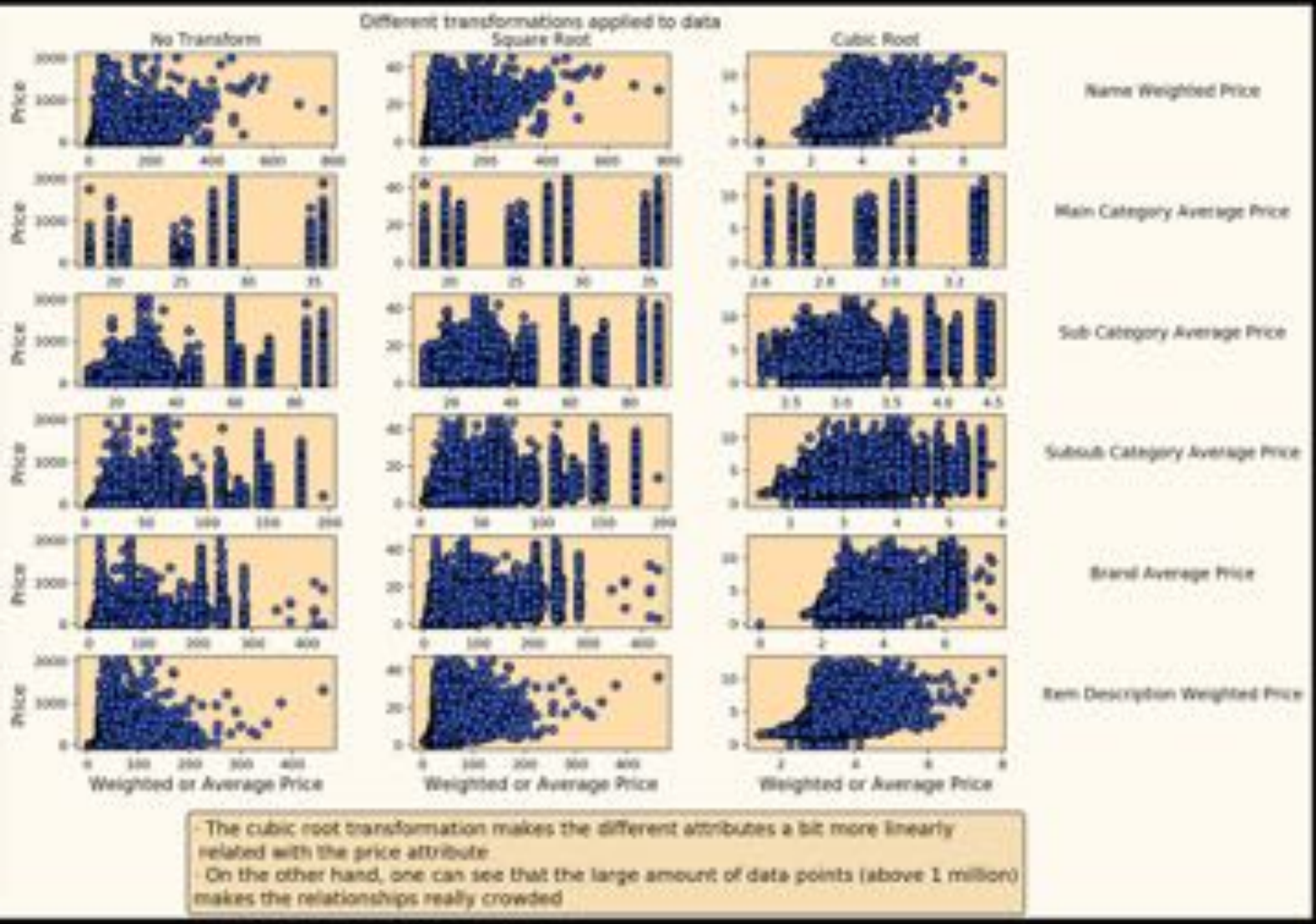
- Further considering the shipping fee of items, there might be a relation between who pays the shipping fee and the average price of products

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

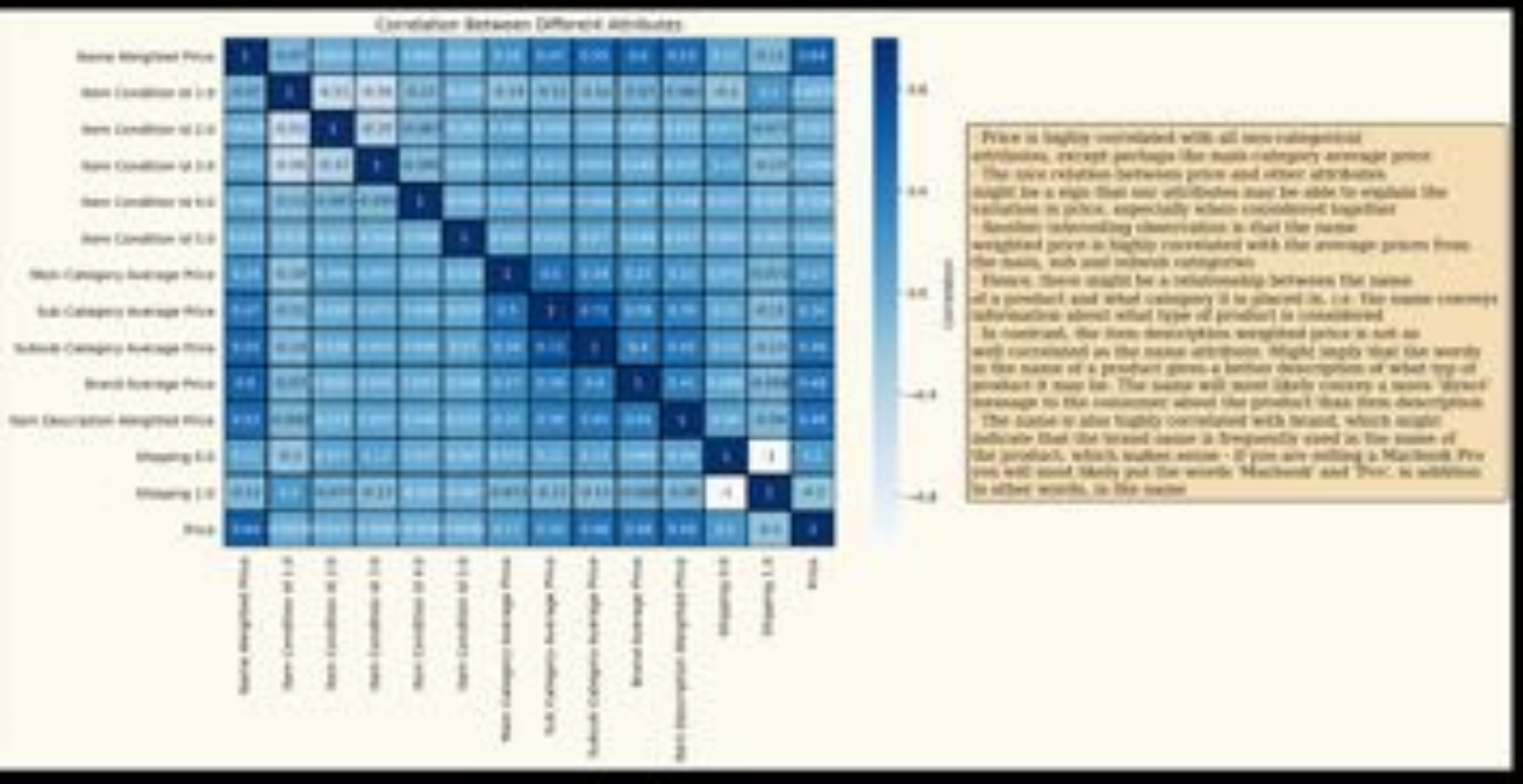


Mercari Price Analysis - Prediction Analysis - Transformation of variables, correlation among variables

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

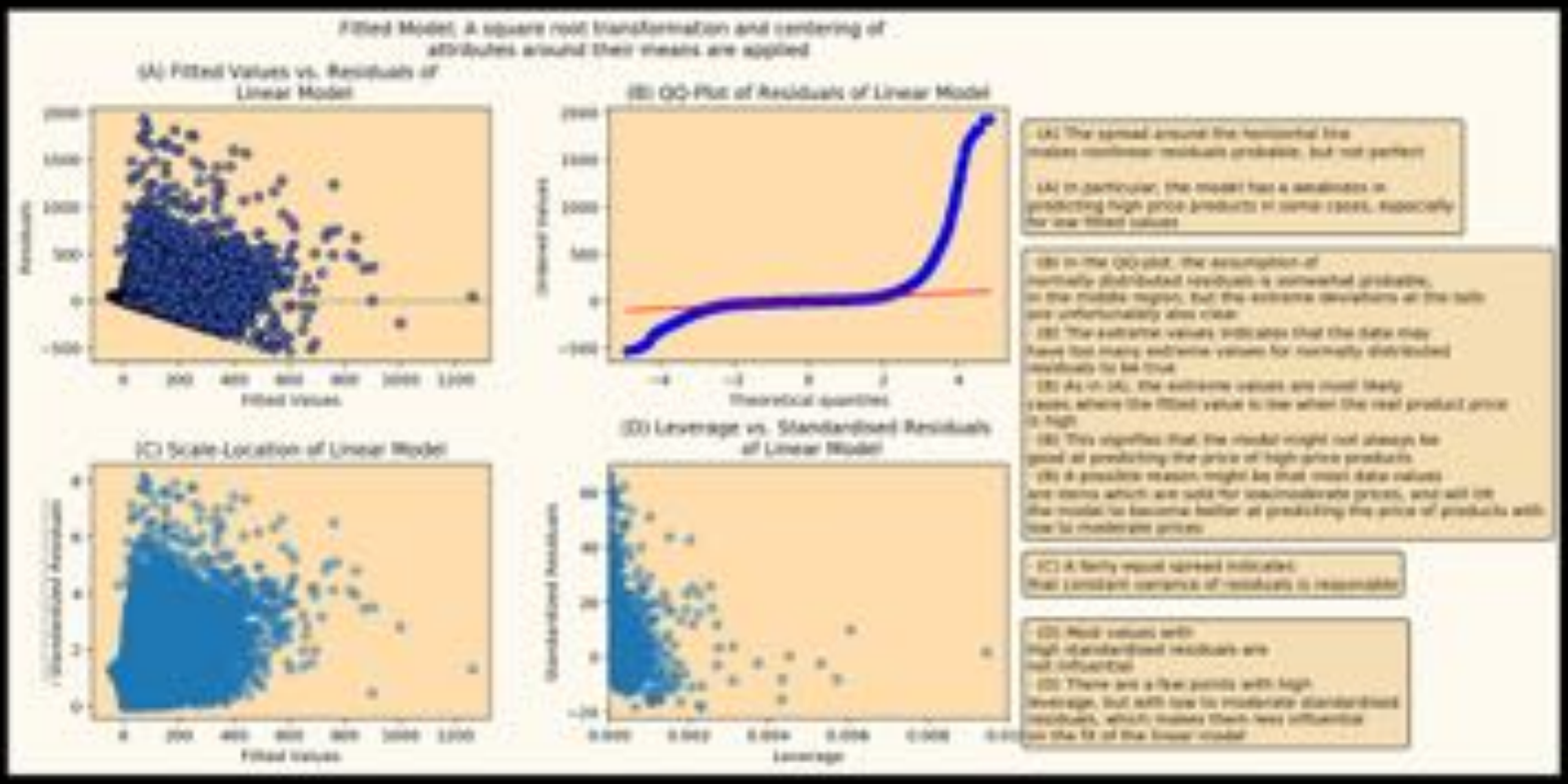


- A key idea in linear models is to analyze the correlation among all variables considered, especially the correlations with respect to the price variable

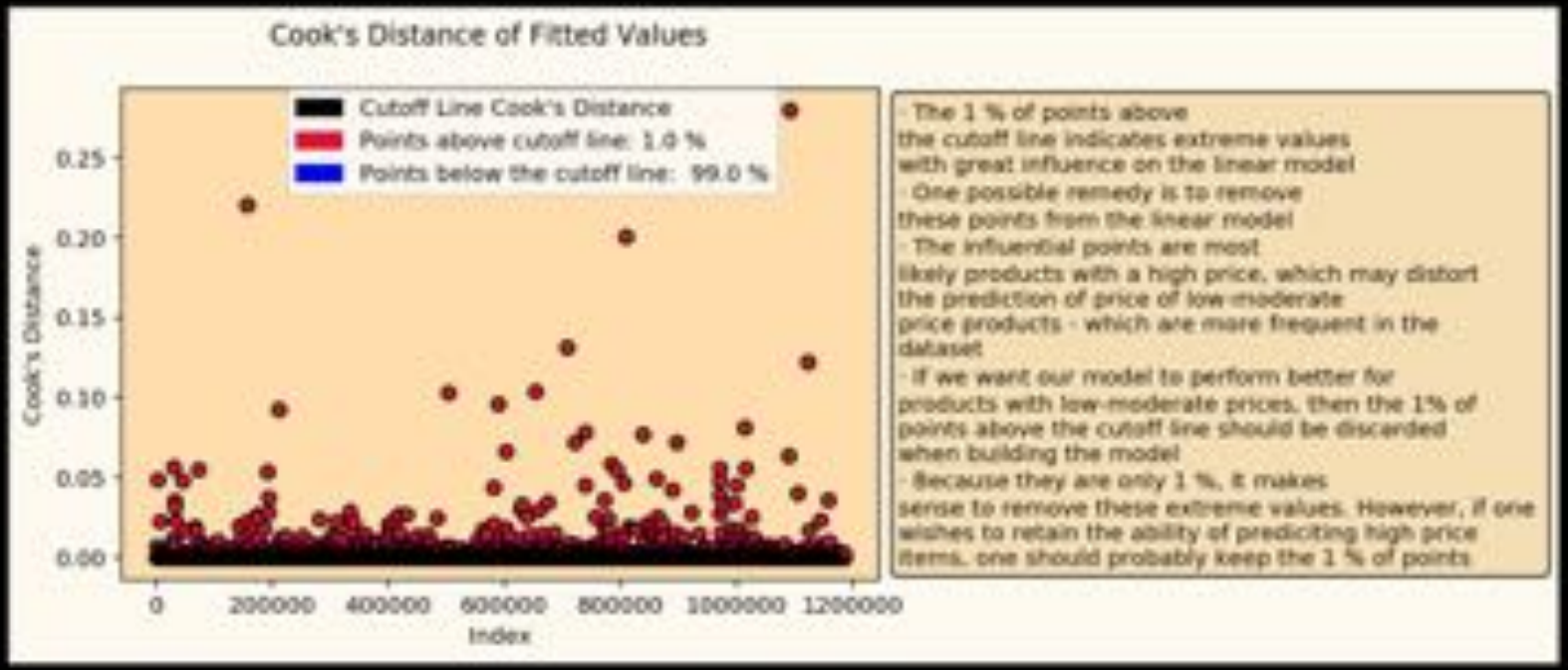


Mercari Price Analysis - Prediction Analysis - First linear model; Diagnostics and Cook's Distance

- 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

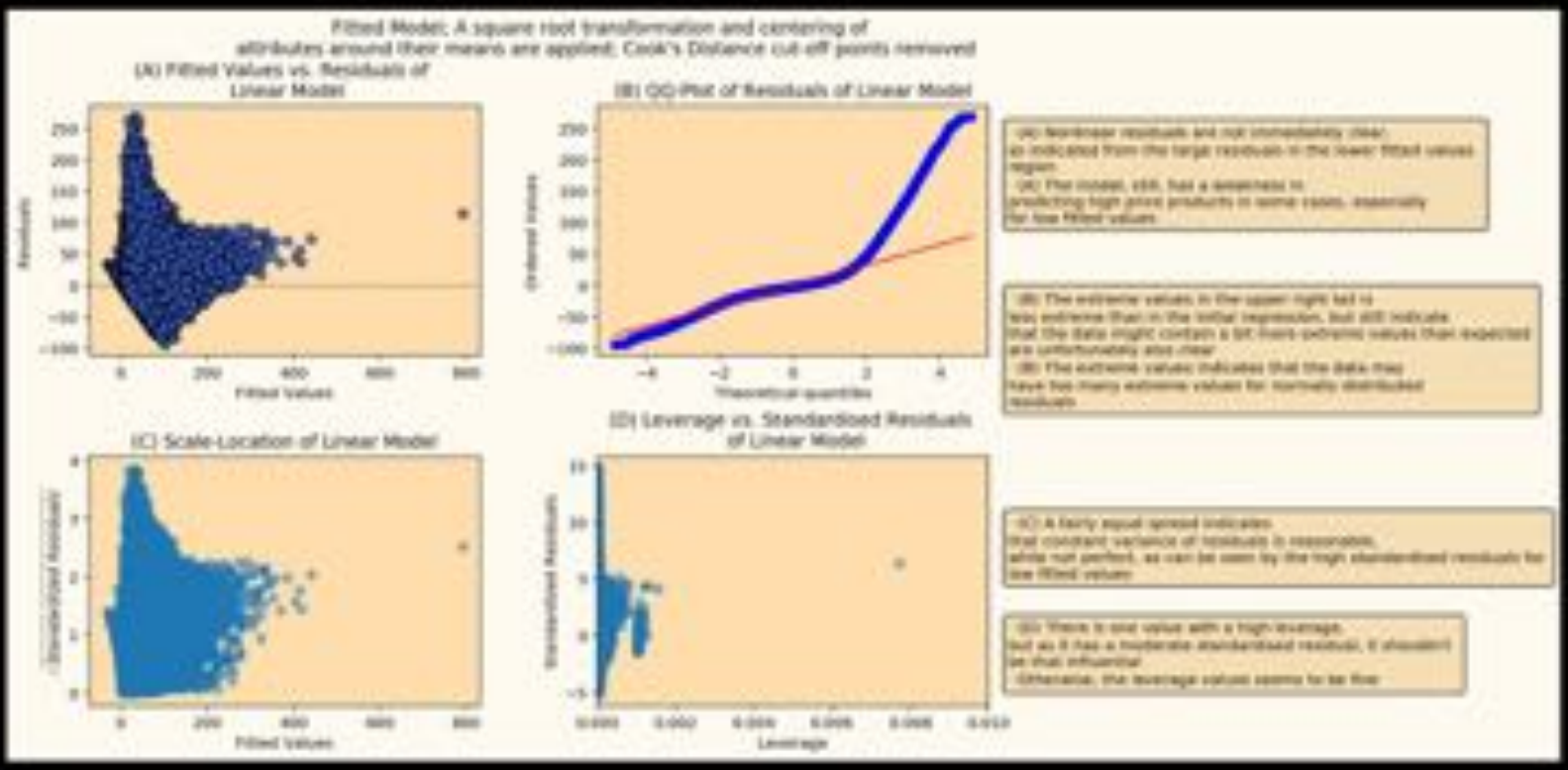


- Further describing the linear model, especially illustrating data points with influence on the fit of the model, the Cook's Distance can be considered

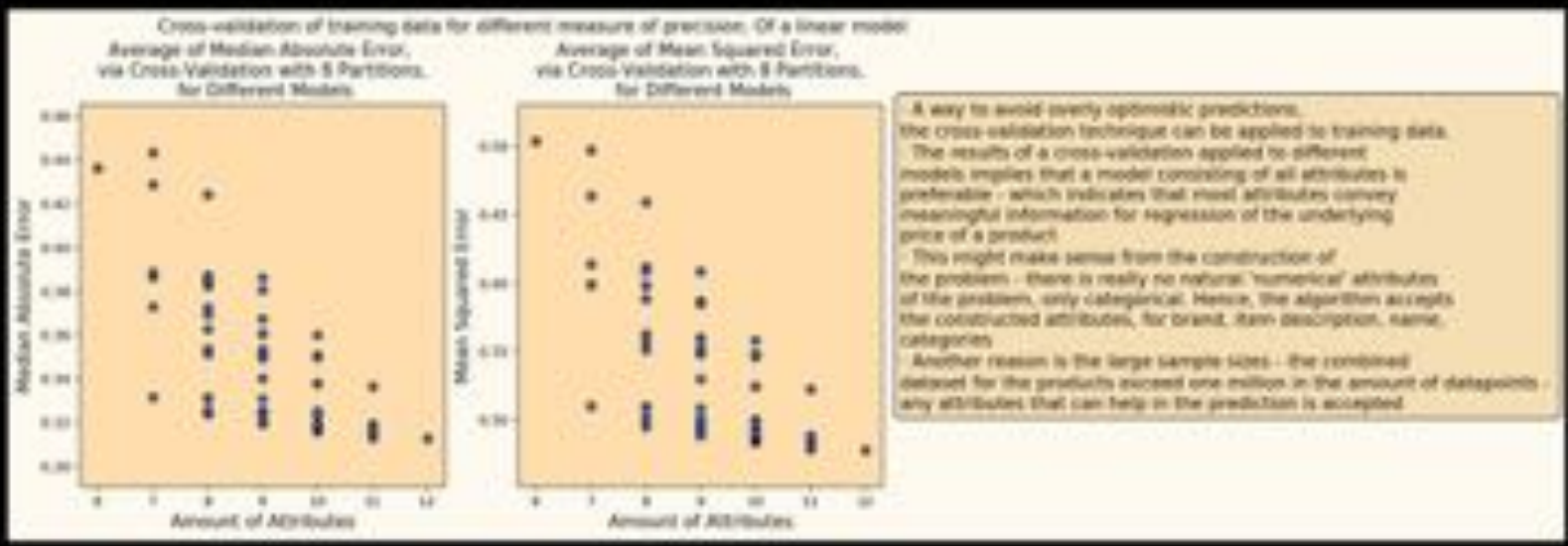


Mercari Price Analysis - Prediction Analysis - Linear model after Cook's Distance analysis, and different performance measures to choose an optimal model

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

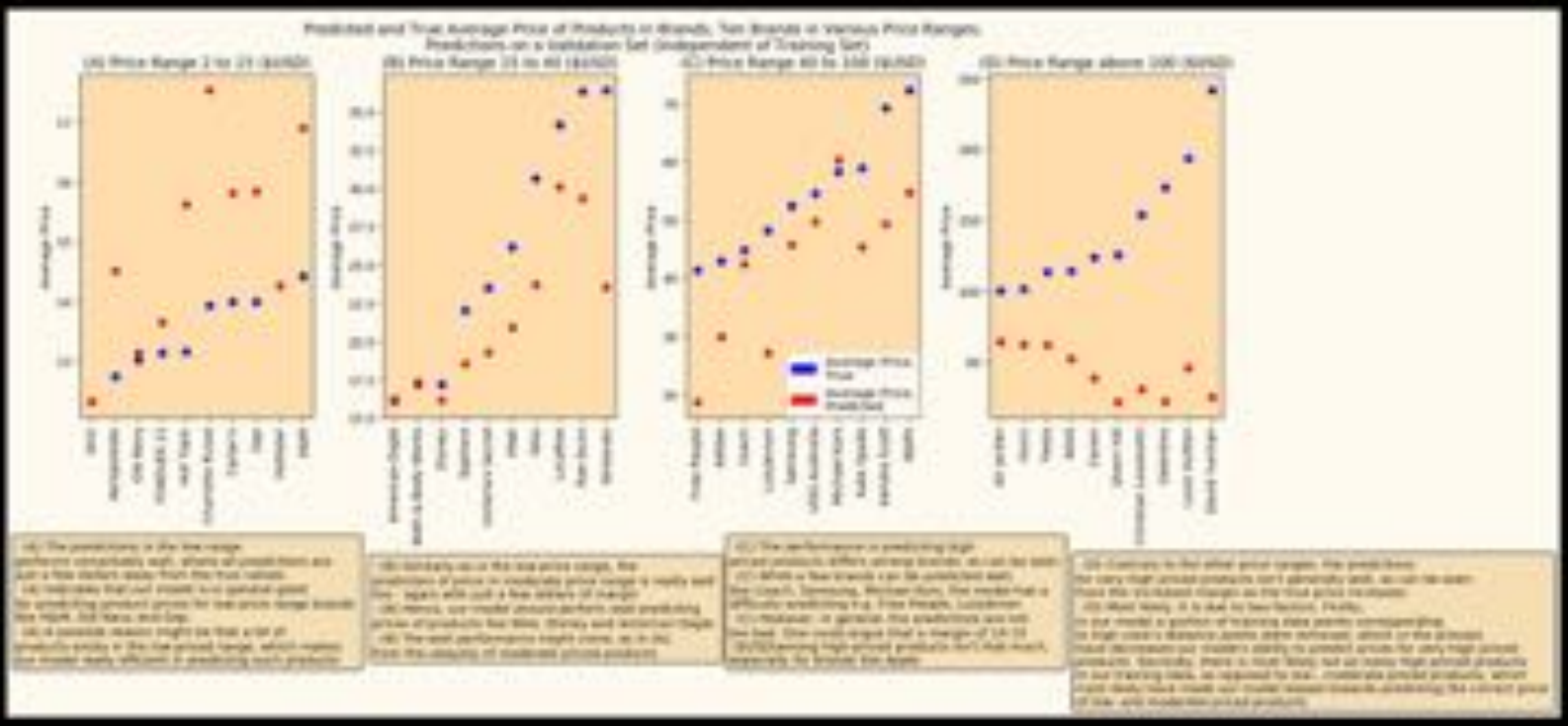


- The aim is to find an optimal model. For this, we consider all combinations of attributes (where all binary attributes are kept in the model), and for each combination we evaluate two performance measures. Lastly, based on the performances, we pick a final, optimal model

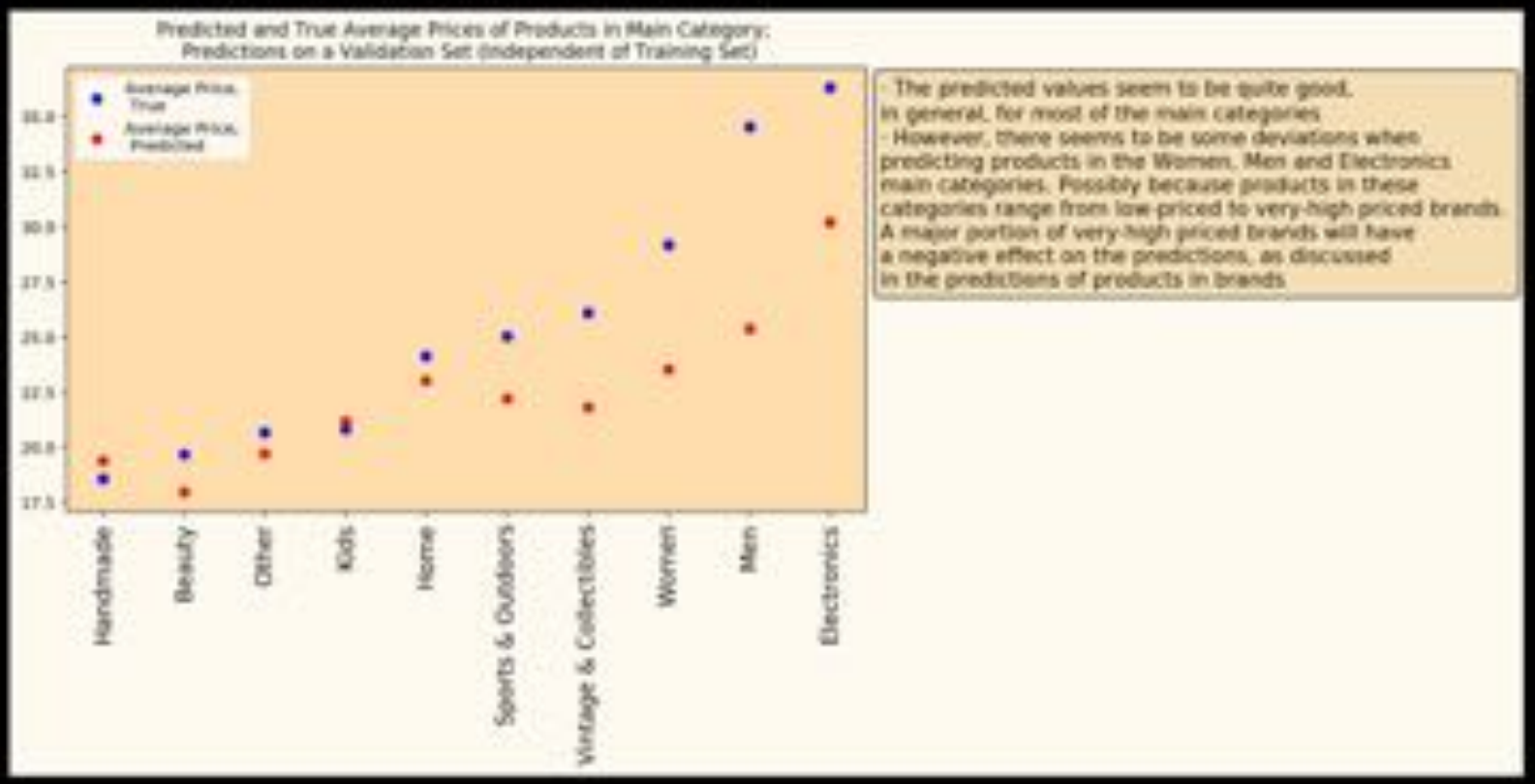


Mercari Price Analysis - Prediction Analysis - Optimal model in predicting prices of products in brands, and predicting products in different main categories

- 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



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



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