# Capstone Project 2: Price Prediction of an online marketplace

In this project, the goal is to predict the price of items that are listed on Mercari marketplace so that when a seller attempts to place their items on the website, they get a suggested price automatically. In the first phase of the project, we start by looking into the data sets we are given. The sets are from [www.kaggle.com](http://www.kaggle.com/) and the sets are found here: <https://www.kaggle.com/saitosean/mercari/version/1>

Since the data set is from Kaggle, the test set does not have the price. Therefore, we will only work with the train set and in the Machine Learning phase, we will split the train set into test and train.

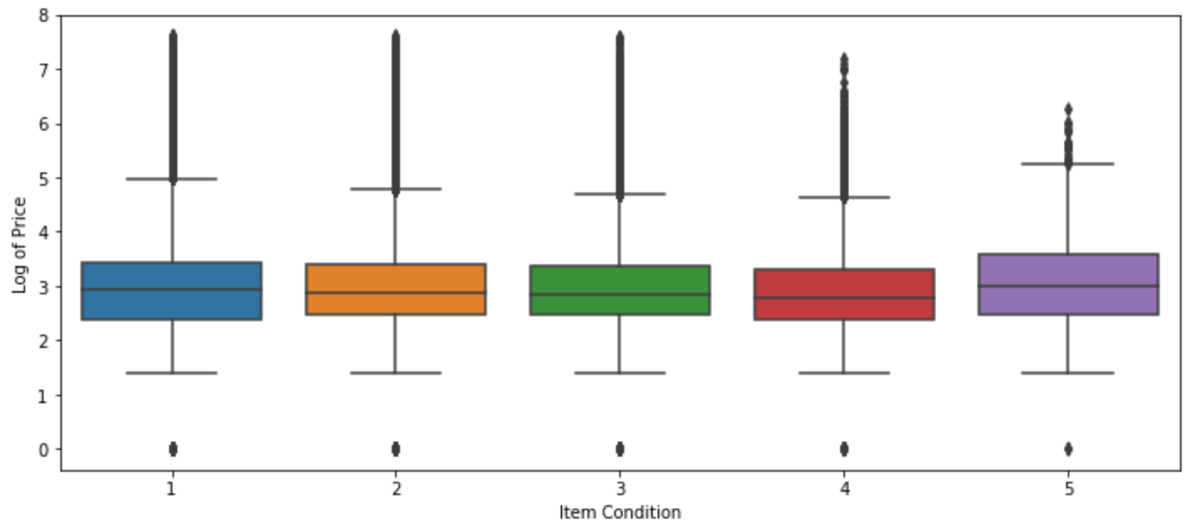
The following is the name of the attributes in our data set:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| test\_id | the id of the listing |
| name | the title of the listing as it appears on the item profile |
| item\_condition\_id | the condition of the items provided by the sellers; 1 is New With Tags, all the way to 5 as the condition gets worse |
| category\_name | category of the listing (we will split up the category name in this project to create more similarities in seemingly different categories) |
| brand\_name | the name of the brand for the item on sale |
| shipping | 1 if shipping fee is paid by seller and 0 if shipping is paid by buyer |
| item\_description | the full description of the item as it appears on the item profile |
| price | the price that the item was sold for. This is target variable that we will predict in this study |

The target variable is `price` which we intend to predict based on the other variables.

**Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?**

As we explored the data in the EDA phase, the `item\_condition` attribute and `price` is not really correlated, meaning that different `item\_condition`’s have almost the same average price when investigating the box plot.



### In the EDA phase of the project, we also graphed the distribution of `price` based on the types of `shipping`’s, namely `shipping` by the seller or the buyer. The distribution of the `price` was more or less the same and it seems like, overall, the price of the items when it's free shipping is less expensive. Namely, more expensive items are paid by the buyer but the shift between the two was not a noticeable shift.

