Mercari Price Suggestion Challenge

Ali Ezzat, Mohammed Ali

December 30, 2017

Contents

About **1**

Exploration **1**
Data Loading 1
Feature relationships 2

Wrangling **4**

Analysis **4**

Model Building **4**

Conclusion **4**

About

Mercari's challenge is to build an algorithm that automatically suggests the right product prices. You'll be provided user-inputted text descriptions of their products, including details like product category name, brand name, and item condition.

Exploration

Data Loading

- Load training data

```
train <- read.csv("Mercari-Price-Suggestion-Challenge/data/train.tsv", sep = "\t",
    stringsAsFactors = FALSE)
```

- Inspect structure

```
glimpse(train)
```

```
## Observations: 593,376
## Variables: 8
## $ train_id          <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13...
## $ name              <chr> "MLB Cincinnati Reds T Shirt Size XL", "Raze...
## $ item_condition_id <int> 3, 3, 1, 1, 1, 3, 3, 3, 3, 3, 2, 1, 2, 1, 3,...
## $ category_name     <chr> "Men/Tops/T-shirts", "Electronics/Computers ...
## $ brand_name        <chr> "", "Razer", "Target", "", "", "", "Acacia S...
## $ price             <dbl> 10, 52, 10, 35, 44, 59, 64, 6, 19, 8, 8, 34,...
## $ shipping          <int> 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0,...
## $ item_description  <chr> "No description yet", "This keyboard is in g...
```

- All about factors It seems there are some columns are factors, let us convert them to the right format

```
#item condition factor
train$item_condition_id <- as.factor(train$item_condition_id)
levels(train$item_condition_id)
```

```
## [1] "1" "2" "3" "4" "5"
```

```
table(train$item_condition_id)
```

```
##
##      1      2      3      4      5
## 256121 150564 172980  12738    973
```

```
#shupping
train$shipping <- as.factor(train$shipping)
levels(train$shipping)
```

```
## [1] "0" "1"
```

```
table(train$shipping)
```

```
##
##      0      1
## 328556 264820
```

- Price summary

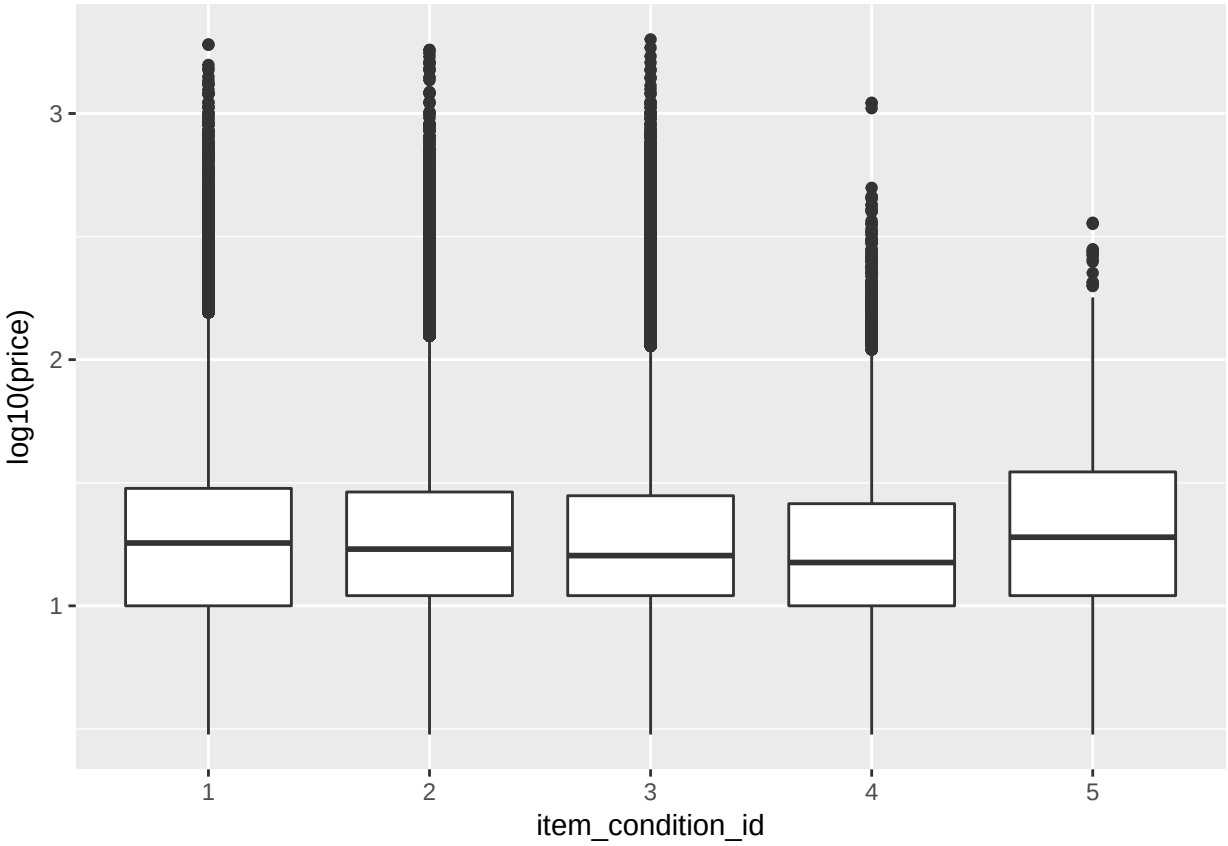
```
summary(train$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   17.00   26.69   29.00 2000.00
```

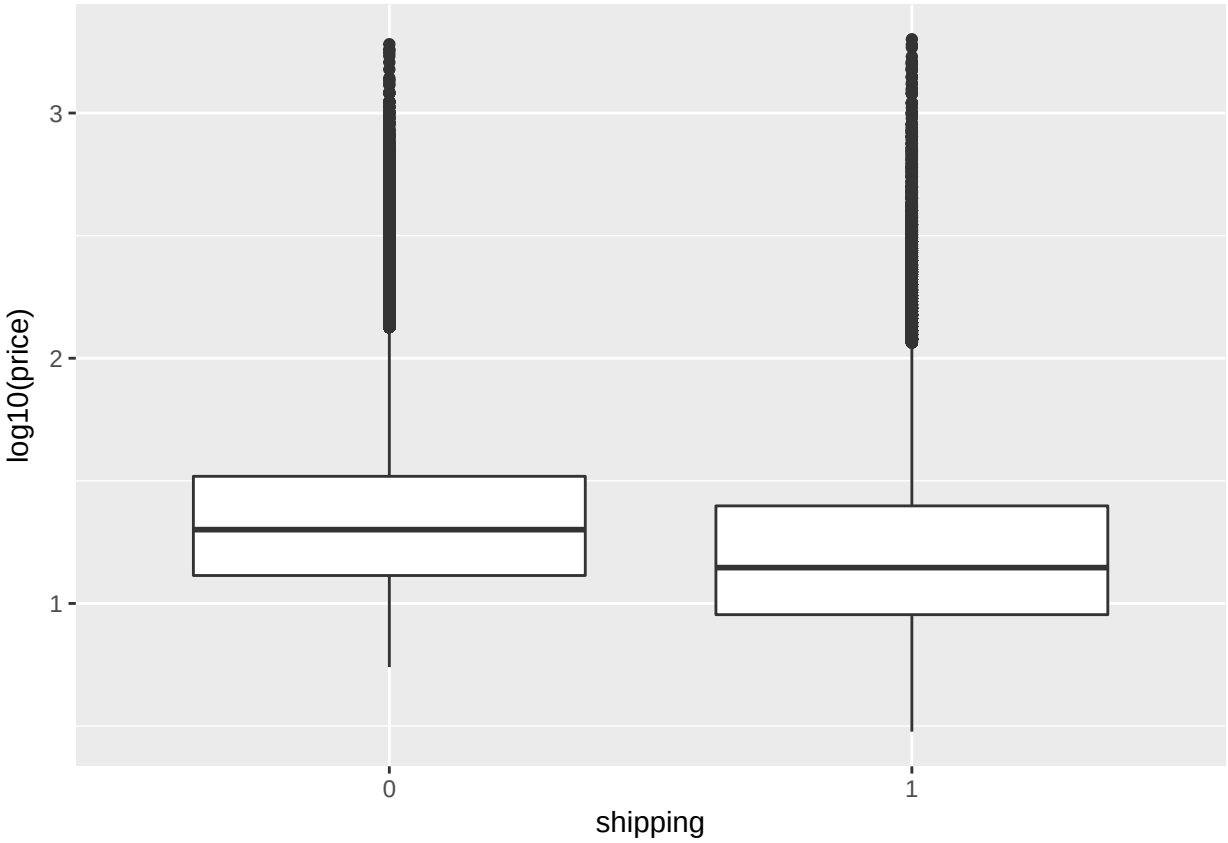
Feature relationships

```
ggplot(train, aes(x = item_condition_id, y = log10(price))) +
  geom_boxplot()
```

```
## Warning: Removed 311 rows containing non-finite values (stat_boxplot).
```




```
ggplot(train, aes(x = shipping, y = log10(price))) +
  geom_boxplot()
```



Wrangling

Analysis

Model Building

Conclusion