Analysis of The IKEA Furniture Price

Group 21-Ananth Padakannaya,Nivedita Patil,Li Wang,Wanqing Yang,Boyao Ma

06/07/2021

1 Introduction

Data set provided is from Ikea (Saudi Arabia), It is of interest to determine which properties of a furniture determine where the price is greater than 1000 riyals.

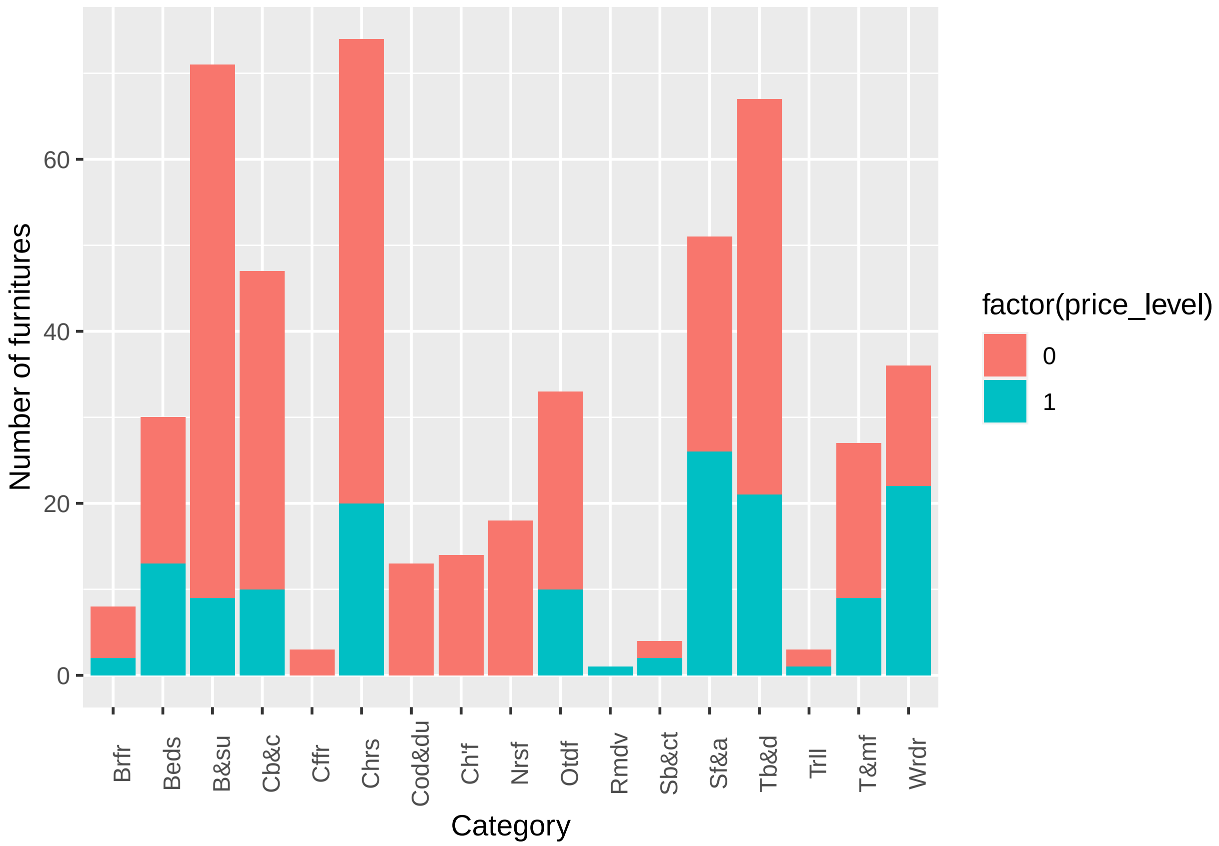
2 Exploratory Data Analysis

Table 1: Summary statistics for observations with chosen variables.

	price	sellable_online	other_colors	depth	height	width	price_level
	Min. : 3.0	Mode :logical	Length:500	Min. : 1.00	Min. : 3.0	Min. : 2.0	Min. :0.000
	1st Qu.: 168.8	FALSE:1	Class :character	1st Qu.: 37.00	1st Qu.: 68.0	1st Qu.: 56.0	1st Qu.:0.000
	Median : 457.0	TRUE :499	Mode :character	Median : 46.00	Median : 83.0	Median : 80.0	Median :0.000
	Mean : 991.1	NA	NA	Mean : 53.34	Mean :102.3	Mean :101.1	Mean :0.292
	3rd Qu.:1245.0	NA	NA	3rd Qu.: 60.00	3rd Qu.:123.8	3rd Qu.:134.2	3rd Qu.:1.000
	Max. :8551.0	NA	NA	Max. :252.00	Max. :251.0	Max. :367.0	Max. :1.000
	NA	NA	NA	NA's :191	NA's :146	NA's :80	NA

We first took 1000 as a dividing point according to the problem, and added a new list of binary variables named price_level. Furniture with a price greater than 1000 takes 1, otherwise it takes 0. Then we performed descriptive statistical analysis based on these selected variables. . . .(Then write some analysis)

3 Visualization of the data



4 Formal Data Analysis

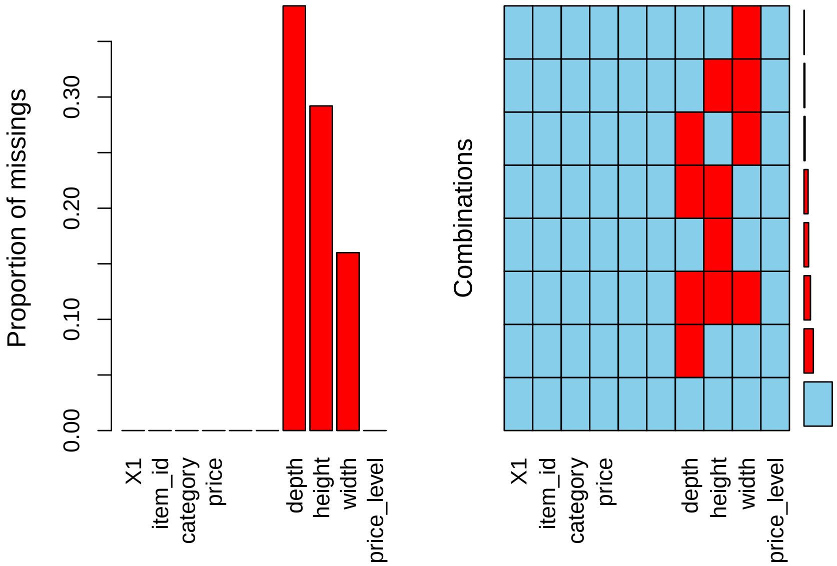


Figure 1: Missing original data.

Through the above figure, we found that there are many missing values and the missing data is mainly concentrated in three explanatory variables, namely depth, length and width. And the three horizontal red squares indicate that these three data are missing at the same time. If we ignore or delete these missing data directly, it will have a great impact on the analysis of the data. So we have to use multiple imputation to fill in missing data.

```
 iter imp variable
  1   1  depth  height  width
  1   2  depth  height  width
  1   3  depth  height  width
  1   4  depth  height  width
  1   5  depth  height  width
  2   1  depth  height  width
  2   2  depth  height  width
  2   3  depth  height  width
  2   4  depth  height  width
  2   5  depth  height  width
  3   1  depth  height  width
  3   2  depth  height  width
  3   3  depth  height  width
  3   4  depth  height  width
  3   5  depth  height  width
  4   1  depth  height  width
  4   2  depth  height  width
  4   3  depth  height  width
  4   4  depth  height  width
  4   5  depth  height  width
  5    1   depth  height  width
  5    2   depth  height  width
  5    3   depth  height  width
  5    4   depth  height  width
  5    5   depth  height  width
```

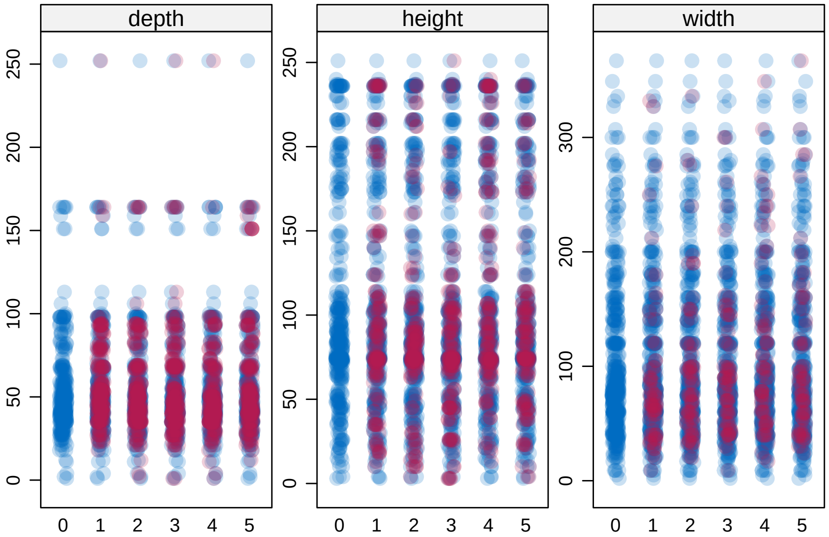


Figure 2: Data situation of multiple imputation method.

```
iter imp variable
 1    1   depth  height  width
 1    2   depth  height  width
 1    3   depth  height  width
 1    4   depth  height  width
 1    5   depth  height  width
 2    1   depth  height  width
 2    2   depth  height  width
 2    3   depth  height  width
 2    4   depth  height  width
 2    5   depth  height  width
 3    1   depth  height  width
 3    2   depth  height  width
 3    3   depth  height  width
 3    4   depth  height  width
 3    5   depth  height  width
 4    1   depth  height  width
 4    2   depth  height  width
 4    3   depth  height  width
 4    4   depth  height  width
 4    5   depth  height  width
 5    1   depth  height  width
 5    2   depth  height  width
 5    3   depth  height  width
```

```
  5   4  depth  height  width
  5   5  depth  height  width
```

According to the picture, we can view the data interpolation. The blue point is the original data, and the red point is the interpolation data. We can see that the two color points are relatively overlapped, indicating that the interpolation is very good. Then, we use the interpolated data to fit the generalized linear model.

Table 2: The coefficient table of the first model.

term	estimate	std.error	statistic	df	p.value	2.5 %	97.5 %
(Intercept)	-0.0363532	0.4389662	-0.0828155	24.68444	0.9346662	-0.9410074	0.8683010
sellable_onlineTRUE	-0.3755903	0.4219650	-0.8900982	30.31691	0.3804224	-1.2369803	0.4857996
other_colorsYes	0.0206518	0.0340456	0.6065940	139.22004	0.5451083	-0.0466614	0.0879651
depth	0.0048354	0.0007937	6.0921034	13.57657	0.0000317	0.0031281	0.0065428
height	0.0015915	0.0003199	4.9742673	60.31613	0.0000058	0.0009516	0.0022314
width	0.0027384	0.0003230	8.4790438	37.08809	0.0000000	0.0020841	0.0033928

We use price_level as the response variable. Because it is a binary variable, so we can use a logistic regression model for the probability of whether the price is greater than 1000. Through the above table, we found that the P values of the two categorical variables(sellable_online and other_colors) are both greater than 0.05 and their confidence interval contains 0, so it means that these two items are not significant in this model, and we need to eliminate these two variables. Next, we use the remaining variables to perform a new modeling.

$$log(\frac{\widehat{\text{p}}_{\text{i}}}{1-\widehat{\text{p}}_{\text{i}}}) = \widehat{\alpha} + \widehat{\beta} * \text{depth}_{\text{i}} + \widehat{\gamma} * \text{height}_{\text{i}} + \widehat{\delta} * \text{width}_{\text{i}}$$

where

- the $\widehat{\text{p}}_{\text{i}}$: the probability of whether the price is greater than 1000 for the ith furniture.
- the $\widehat{\alpha}$: the intercept of the regression line.
- the $\widehat{\beta}$: the coefficient for the first explanatory variable depth.
- the $\widehat{\gamma}$: the coefficient for the second explanatory variable height.
- the $\widehat{\delta}$: the coefficient for the second explanatory variable width.

When this model is fitted to the data, the following estimates of α (intercept) and β,γ and δ are returned:

Table 3: The coefficient table of the final model.

term	estimate	std.error	statistic	df	p.value	2.5 %	97.5 %
(Intercept)	-0.4065488	0.0484309	-8.394417	45.71185	0.00e+00	-0.5040517	-0.3090460
depth	0.0048464	0.0007943	6.101855	13.39543	3.31e-05	0.0031357	0.0065571
height	0.0015832	0.0003177	4.983428	62.64053	5.20e-06	0.0009483	0.0022182
width	0.0027766	0.0003187	8.711938	35.64118	0.00e+00	0.0021300	0.0034231

According to the coefficients in the above table, we can get the final model as follows:

$$log(\frac{\widehat{\text{p}}_{\text{i}}}{1-\widehat{\text{p}}_{\text{i}}}) = -0.4065 + 0.0048 * \text{depth}_{\text{i}} + 0.0016 * \text{height}_{\text{i}} + 0.0028 * \text{width}_{\text{i}}$$

This is equivalent to:

$$\widehat{\text{p}}_{\text{i}} = \frac{exp(-0.4065 + 0.0048 * \text{depth}_{\text{i}} + 0.0016 * \text{height}_{\text{i}} + 0.0028 * \text{width}_{\text{i}})}{1 + exp(-0.4065 + 0.0048 * \text{depth}_{\text{i}} + 0.0016 * \text{height}_{\text{i}} + 0.0028 * \text{width}_{\text{i}})}$$

Lily(write something to explain this formula)

5 Conclusions and Future Works

6 References