Analysis of The IKEA Furniture Price

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

06/07/2021

1 Introduction

The IKEA store, known for its stylish design and affordable price for young people, has become one of the most popular and recognized furniture retailers in the global marketplace. The company has expanded its global reach very quickly. From Wikipedia, we know that in the early, the first IKEA store in Saudi Arabia opened in 1983 and three stores are opening in this region now. In this study, it is of interest to explore which properties of furniture influence whether they cost more than 1000 Saudi Riyals. We use data from IKEA (Saudi Arabia), including 500 observations and the variables of category, price, sellable_online, other_colors, depth, height and width. In particular, this report further explores the impact level of depth, height and width on the price.

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 of 500 observations.

Table 1 shows that the price of the furniture ranges from 3.0 to 8551.0, with the middle 50% falling between 168.8 and 1245.0 and an average price of 991.1. We can also observe that the middle 50% of depth is between 37.0 and 60.0, with depth of 53.3 on average. Similarly, the middle 50% of height lies between 68.0 and 123.8, with an average value of 102.3. It also shows the mean value of width is 101.1 and the middle 50% falls between 56.0 and 134.2. Besides, we may look at the sellable_online variable is a logical variable, with 499 TRUE values demonstrating that these quantities of items are available to purchase online and 1 FALSE value demonstrating that 1 item is unavailable. In terms of other_colors, it shows that this is a character variable.

3 Visualization of the data

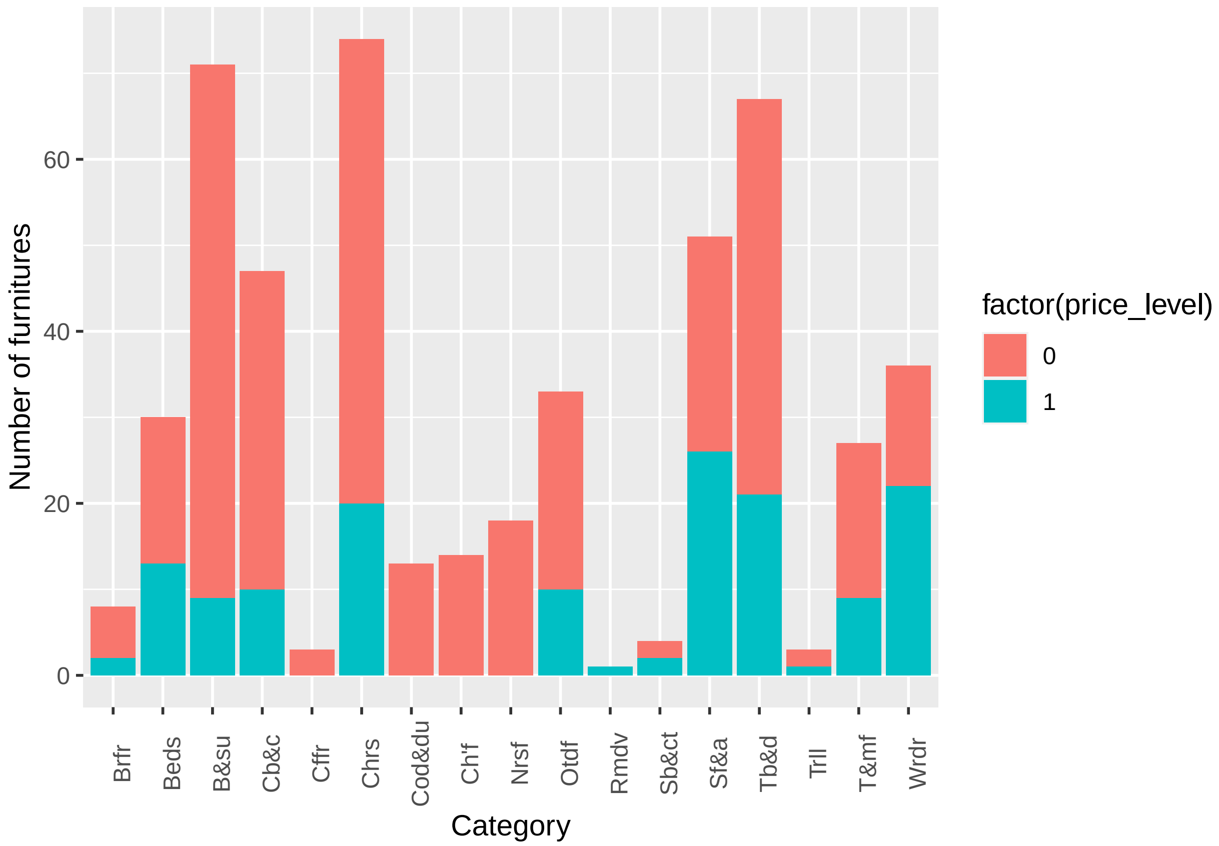

Figure 1 shows the number of furniture based on categories with a price_level 1 (greater than 1000) and 0 (less than 1000). It is obvious that the number of furniture whose price is greater than 1000 is far less than that of furniture whose price is less than 1000 and the prices of furniture in certain categories are all under 1000, such as Cffr, Cod&du, Ch'f and Nrsf. In addition, the largest difference is found in the quantity of B&su costing more than 1000 and less than 1000.

4 Formal Data Analysis

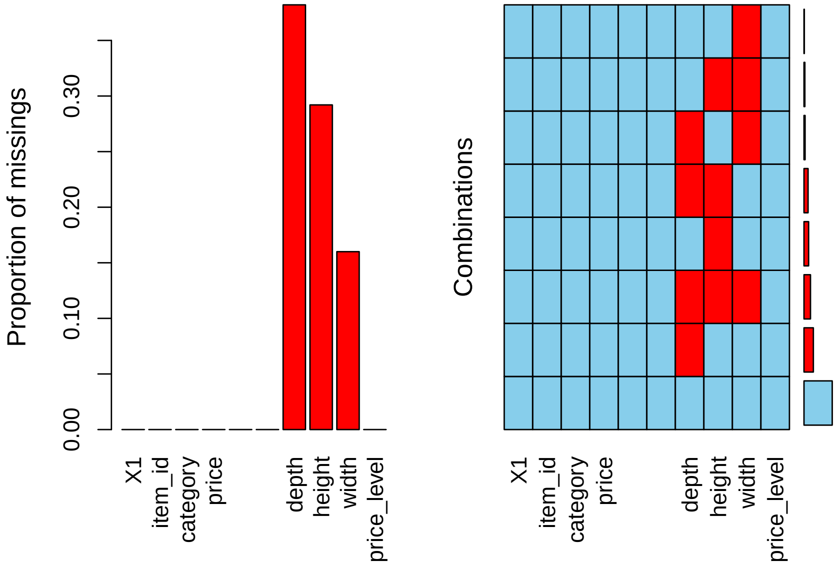

Figure 1: Missing original data.

Through the above figure, we found anomalies 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
```

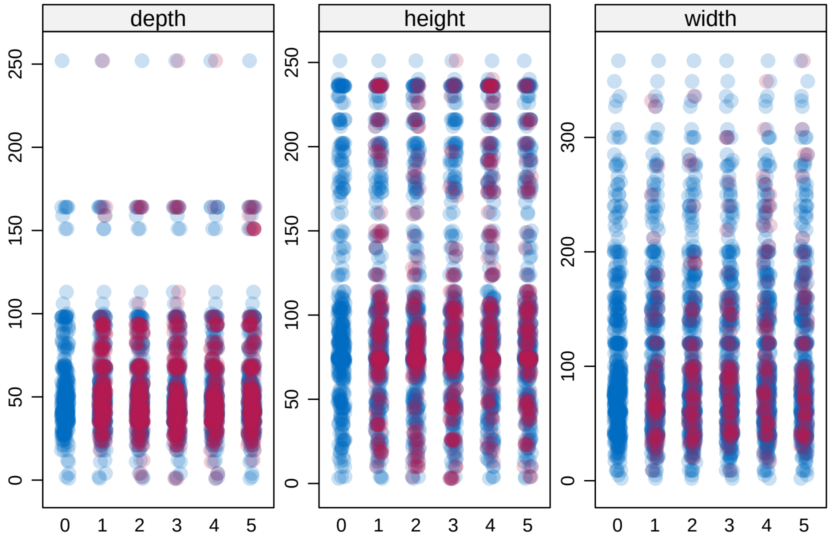

Figure 2: Data situation of multiple imputation method.

```
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 chose the fourth database of multiple imputation for generalized linear model analysis.

```
Call:
glm(formula = price_level ~ sellable_online + other_colors +
    depth + height + width, family = binomial(link = "logit"),
    data = ikea)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8623  -0.4960  -0.2712   0.1969   2.5461

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           6.424137 882.743717   0.007    0.994
sellable_onlineTRUE -13.870206 882.743490  -0.016    0.987
other_colorsYes       0.231993   0.290954   0.797    0.425
depth                 0.047421   0.007090   6.689 2.25e-11 ***
height                0.012512   0.002470   5.066 4.07e-07 ***
width                 0.022326   0.002864   7.796 6.40e-15 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 603.93  on 499  degrees of freedom
Residual deviance: 323.35  on 494  degrees of freedom
AIC: 335.35

Number of Fisher Scoring iterations: 13
```

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, 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.

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

where

- the $\widehat{\mathrm{p}}_{\mathrm{i}}$: 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:

```
Call:
glm(formula = price_level ~ depth + height + width, family = binomial(link = "logit"),
    data = ikea)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8754  -0.4854  -0.2761   0.2018   2.5107

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.342133   0.654665 -11.215  < 2e-16 ***
depth        0.046946   0.007027   6.681 2.37e-11 ***
height       0.012162   0.002447   4.969 6.73e-07 ***
width        0.022909   0.002833   8.086 6.19e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 603.93  on 499  degrees of freedom
Residual deviance: 324.69  on 496  degrees of freedom
AIC: 332.69

Number of Fisher Scoring iterations: 6
```

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

$$log(\frac{\widehat{\mathrm{p}}_\mathrm{i}}{1-\widehat{\mathrm{p}}_\mathrm{i}}) = -7.3421 + 0.0469 * \mathrm{depth}_\mathrm{i} + 0.0122 * \mathrm{height}_\mathrm{i} + 0.0229 * \mathrm{width}_\mathrm{i}$$

This is equivalent to:

$$\widehat{\mathrm{p}}_\mathrm{i} = \frac{exp(-7.3421 + 0.0469 * \mathrm{depth}_\mathrm{i} + 0.0122 * \mathrm{height}_\mathrm{i} + 0.0229 * \mathrm{width}_\mathrm{i})}{1 + exp(-7.3421 + 0.0469 * \mathrm{depth}_\mathrm{i} + 0.0122 * \mathrm{height}_\mathrm{i} + 0.0229 * \mathrm{width}_\mathrm{i})}$$

The log-odds of the price of a furniture being greater than 1000 increase by 0.05 for every 1 unit increase in depth when keeping the height and width constant. Similarly, the log-odds increase by 0.01 for every 1 unit increase in height and increases by 0.02 for every unit increase in width.

Table 2: The confidence interval of variables.

	2.5 %	97.5 %
(Intercept)	-8.7113103	-6.1375780
depth	0.0339458	0.0614882
height	0.0074336	0.0170563
width	0.0176286	0.0287658

Table 2 shows the 95% confidence interval for these log-odds, with intercept being of (-8.71, -6.61), depth being of (0.03, 0.06), height being of (0.007, 0.017) and width being of (0.02, 0.03).

Table 3: Odds scale.

	x
(Intercept)	0.0006477
depth	1.0480658
height	1.0122360
width	1.0231734

On the odds scale, the intercept value (0.00006477) gives the probability that the price is greater than 1000 when depth= 0, width=0 and height=0. This is obviously not the feasible range of depth, width and height, so why this value is very close to zero. For depth, there is a probability of 1.05, which means that for each increase in depth by 1 unit, the probability that the furniture price is greater than 1000 increases by 1.06 times. For each unit of the same height, the probability increases by 1.01 times. For each unit increase in width, the probability increases by 1.02 times.

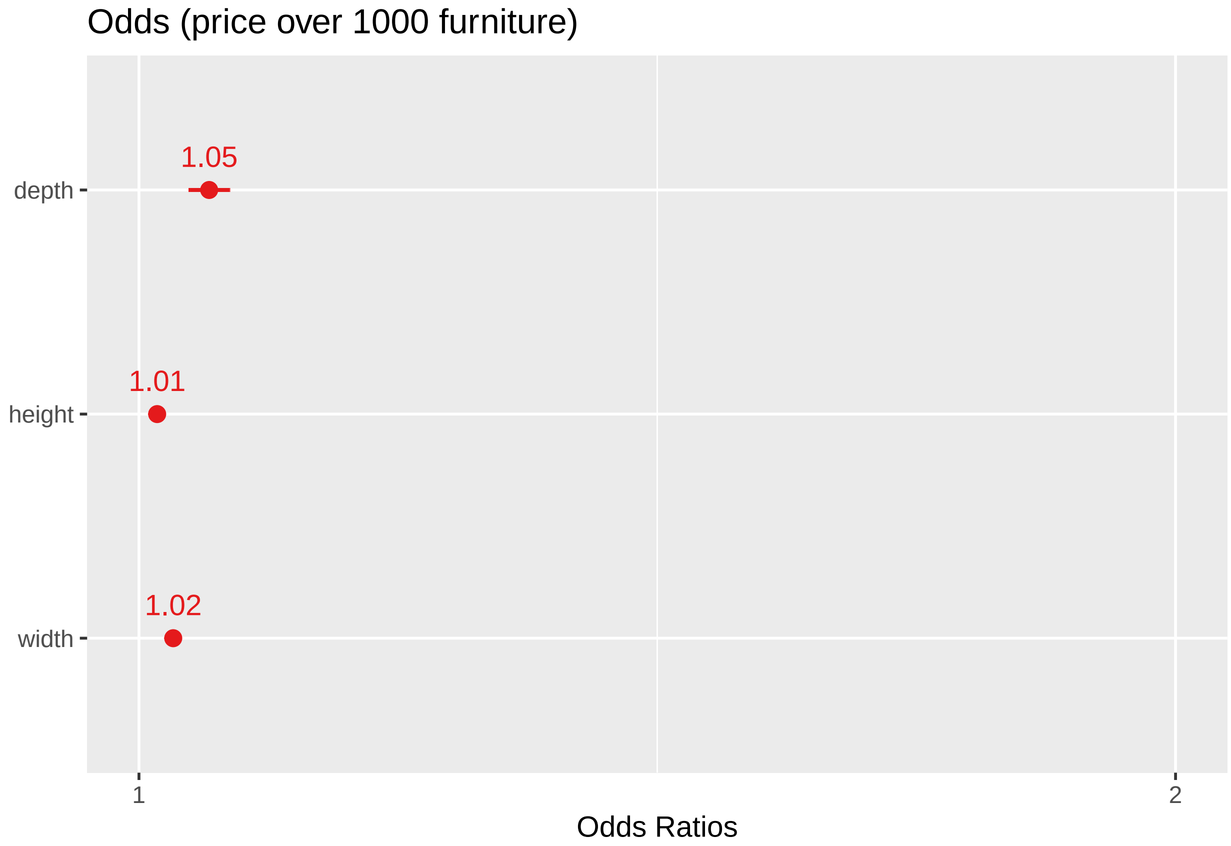

Figure 3: Odds ratios of three explanatory variables.

We can also see the graphic about. . . . . . (lily)

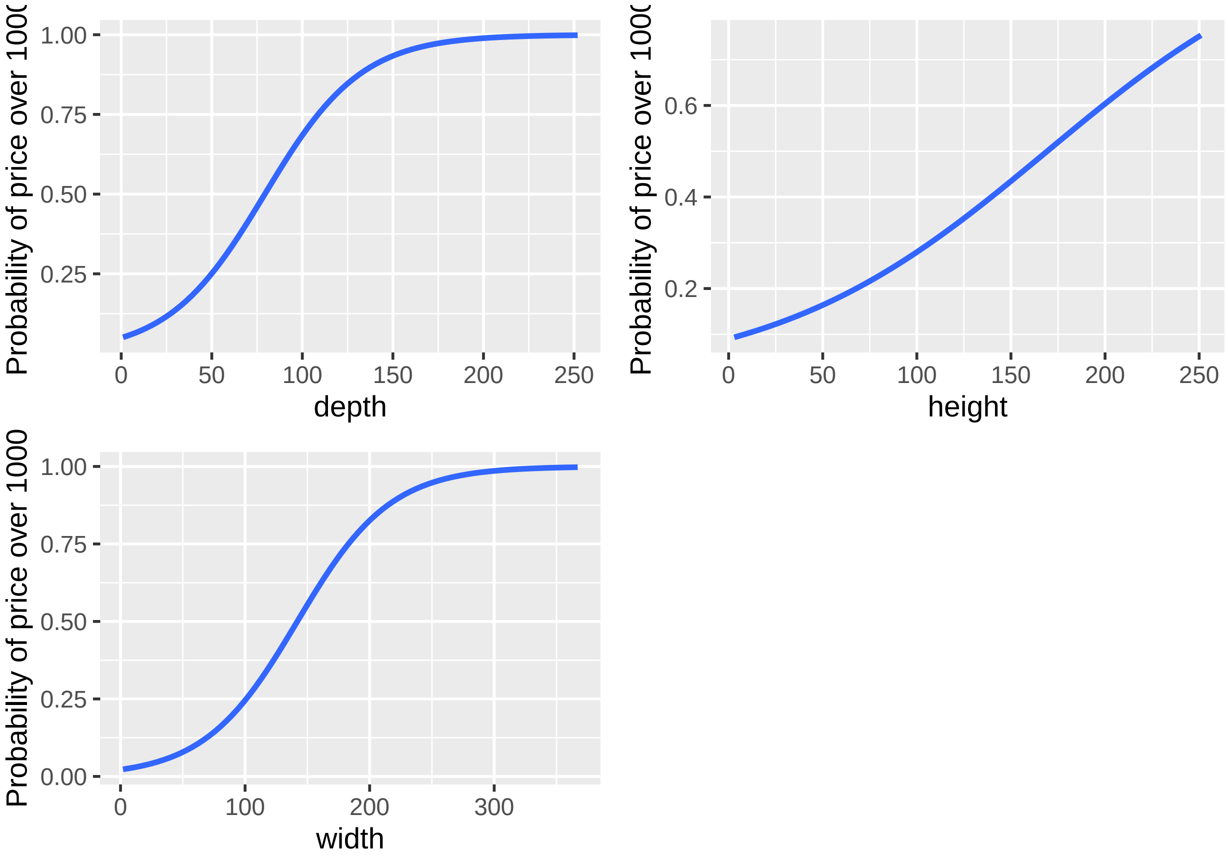

Figure 4: Probability of price over 1000 by three different variables.

lily. . .

5 Goodness of fit

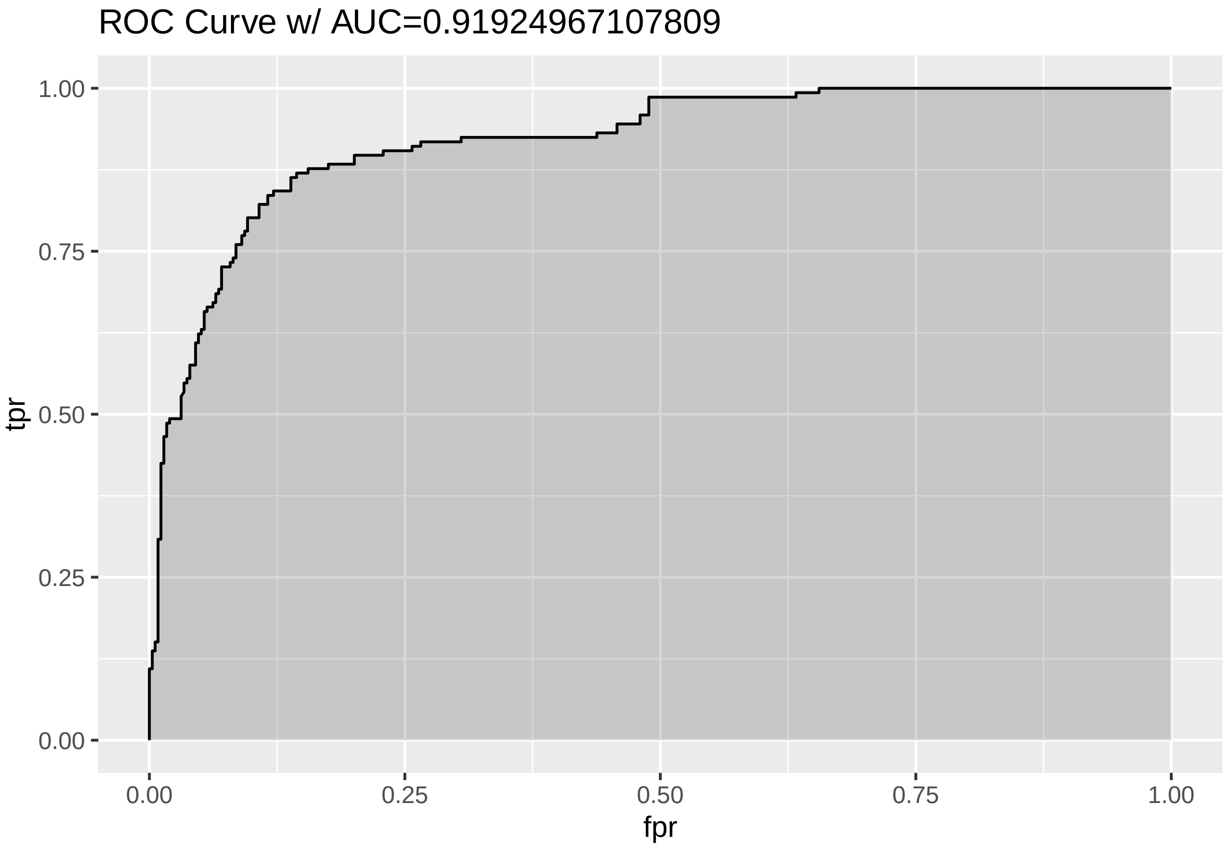

Figure 5: ROC curve.

explain ROC and AUC

6 Conclusions and Future Work

Since the length, width and height of an item are related to the price, we can consider introducing volume variable which is the product of the three as a variable into the logistic regression model in future work.