# DAS2021 - Project 2

Group 22 - Shuang Wu, SHANSHAN LU, Linfeng Guo, Emmanouil Mertzianis, WAN XIE

11/7/2021

```
## Warning: package 'moderndive' was built under R version 4.0.5

## Warning: package 'knitr' was built under R version 4.0.5

## Warning: package 'tidyverse' was built under R version 4.0.5

## Warning: package 'ggplot2' was built under R version 4.0.3

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.4

## Warning: package 'readr' was built under R version 4.0.5

## Warning: package 'purrr' was built under R version 4.0.4

## Warning: package 'dplyr' was built under R version 4.0.4

## Warning: package 'stringr' was built under R version 4.0.5

## Warning: package 'forcats' was built under R version 4.0.5

## Warning: package 'skimr' was built under R version 4.0.5

## Warning: package 'broom' was built under R version 4.0.5

## Warning: package 'ggpubr' was built under R version 4.0.5
```

## 1 Introduction

The objective of this report is to identify the most influential attributes that explain the price of a furnishing product sold by IKEA. To this end, a data set from IKEA Saudi Arabia was collected, containing measurements about 500 items of furniture. The variables considered are:

- `item_id` – unique product ID
- `category` – the furniture category the item belongs to
- `price` – the current price in Saudi Riyals (as recorded on 20/04/2020)

- `sellable_online` – a binary variable to indicate whether the item is available to purchase online
- `other_colors` – a binary variable to indicate whether the item is available in other colours
- `depth` – depth of the item in cm
- `height` – height of the item in cm
- `width` – width of the item in cm

More specifically, we are interested in discovering the most important features out of all available that dictate whether a product is more expensive than 1000 Saudi Riyals. For that reason, we create one more binary variable, `priceMoreThan1000`, to indicate whether a product costs more than 1000 Saudi Riyals.

Throughout the report, we consider various numerical and graphical summaries, followed by the use of an appropriate generalised linear model in order to assess the relationship between the available variables and the 1000 Suadi Riyals threshold related to the price.

# 2 Exploratory Data Analysis

Before conducting any formal data analysis using statistical models, it is useful to explore our data and the relationships between them using numerical and graphical summaries. The following table contains summary statistics for the variables of our data set except for the `item_id` variable as it constitutes an identification variable that does not hold any useful information about the products.

Table 1: Data summary

| | |
|---|---|
| Name | Piped data |
| Number of rows | 500 |
| Number of columns | 8 |
| Column type frequency: | |
| factor | 1 |
| logical | 3 |
| numeric | 4 |
| Group variables | None |

**Variable type: factor**

| skim_variable | n_missing | complete_rate | ordered | n_unique | top_counts |
|---|---|---|---|---|---|
| category | 0 | 1 | FALSE | 17 | Tab: 81, Boo: 78, Cha: 73, Sof: 57 |

**Variable type: logical**

| skim_variable | n_missing | complete_rate | mean | count |
|---|---|---|---|---|
| sellable_online | 0 | 1 | 0.98 | TRU: 492, FAL: 8 |
| other_colors | 0 | 1 | 0.40 | FAL: 302, TRU: 198 |
| priceMoreThan1000 | 0 | 1 | 0.37 | FAL: 317, TRU: 183 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|---|
| price | 0 | 1.00 | 1156.96 | 1471.83 | 5 | 195 | 564 | 1592.5 | 9585 |
| depth | 213 | 0.57 | 54.29 | 31.41 | 3 | 39 | 47 | 58.0 | 210 |
| height | 135 | 0.73 | 103.29 | 61.98 | 1 | 64 | 83 | 145.0 | 321 |
| width | 77 | 0.85 | 106.80 | 75.73 | 3 | 60 | 80 | 141.5 | 387 |

There are several interesting findings from the summary statistics.

Starting from the top of the table, we observe that we have 4 categorical variables (including the newly-created price based one) and 4 numeric variables.

- For the furniture category, we observe no missing values and 17 different categories, while a more thorough view of the different categories follows below.
- The amount of products sold online in our data set completely dominates the amount of the ones that are not, accounting for the 98% of the data set (492 products in total).
- The products available in other colours account for just the 40% of the total amount of products observed. This means that there are only 198 products available in other colours, while 302 products are sold in one colour.
- There are only 183 items priced over 1000 Saudi Riyals, which is roughly one third of the total amount of products observed.
- When the price is considered as a continuous variable, its variance is significantly larger than the rest of the numeric variables, with prices as low as 5 and as high as 9585 Saudi Riyals. However, such a difference can be justified by the different measurement units in which these variables are measured.
- Regarding the depth of the items, it is apparent that almost half of the observed items are lacking a depth measurement (213 items), which is a considerable amount of missing information. Furthermore, comparing it to height and width, the mean value of depth is about half of the corresponding means of the rest of the features. Its standard deviation is about half the standard deviation of height. Moreover, its 50% central sample distribution appears to be lower than that of the others, suggesting that depth is smaller that height and width, in general. Those observations are important since these 3 variables are measured in the same units.
- When it comes to height and width, there are 135 and 77 missing values, respectively. In general, their summary statistics do not present a significant difference. The only two exceptions are their variances and their maximum values. The variance of height is 61.98, while the variance of width is equal to 75.73. Height's maximum value is 321 cm, while width's maximum value is 387 cm.

## 2.1 Missing values

Due to the significant amount of missing information, it is important to address the issue of missing values before proceeding with our data analysis.

Our first step is to remove observations with 2 or more missing values. Such observations can be considered as items with a significant loss of information to the point that they cannot be useful for analysis.

Additionally, in order to avoid removing more information from the data set, we attempt to replace the missing values of the remaining observations. To this end, we calculate the mean of the missing variable using only those observations that belong to the same furniture category as the observation of interest. Essentially, this is equivalent to fitting a linear regression model with the variable of interest as the response and the `category` variable as the predictor and predicting the missing value using the observation's furniture category.

Table 5: Data summary

| | |
|---|---|
| Name | Piped data |
| Number of rows | 383 |
| Number of columns | 8 |
| Column type frequency: | |
| factor | 1 |
| logical | 3 |
| numeric | 4 |
| Group variables | None |

**Variable type: factor**

| skim_variable | n_missing | complete_rate | ordered | n_unique | top_counts |
|---|---|---|---|---|---|
| category | 0 | 1 | FALSE | 17 | Boo: 68, Cha: 46, Tab: 44, Sof: 34 |

**Variable type: logical**

| skim_variable | n_missing | complete_rate | mean | count |
|---|---|---|---|---|
| sellable_online | 0 | 1 | 0.99 | TRU: 380, FAL: 3 |
| other_colors | 0 | 1 | 0.40 | FAL: 228, TRU: 155 |
| priceMoreThan1000 | 0 | 1 | 0.36 | FAL: 246, TRU: 137 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|---|
| price | 0 | 1 | 1172.56 | 1512.96 | 5 | 204.5 | 575 | 1587.5 | 9585 |
| depth | 0 | 1 | 53.65 | 28.48 | 3 | 39.0 | 46 | 60.0 | 210 |
| height | 0 | 1 | 106.28 | 61.17 | 1 | 68.0 | 85 | 147.0 | 321 |
| width | 0 | 1 | 109.66 | 78.87 | 3 | 60.0 | 80 | 150.0 | 387 |

After cleaning the data, the overall results are similar to the ones we had before. However, we point out some key differences and similarities:

- The resulting data set now contains 383 products. This corresponds to the removal of 117 products as these contained 2 or more missing values.
- There is a different ordering of levels when we consider the descending order of `category`'s levels by count. "Tables & desks" was the dominating category before cleaning the data. However, many observations under this category contained 2 or more missing values.
- For the `sellable_online` categorical variable, we observe an even larger proportion of items sold online, accounting for the 99% of the total amount of products.
- For `depth`, `height` and `width`. we observe the same dissimilarities along with slightly wider IQRs for each variable. Also, there is a decrease in the variance of `depth` and an increased variance for `width`.
- Finally, the mean and median values of `price` are increased when considering it as a continuous variable. Also, its variance increased significantly.

## 2.2 Exploring relationships with respect to the price category

The purpose of the report is to identify the most influential variables with respect to the price category of a product. For this purpose, it is useful to explore the relationships between the available variables and the price category.

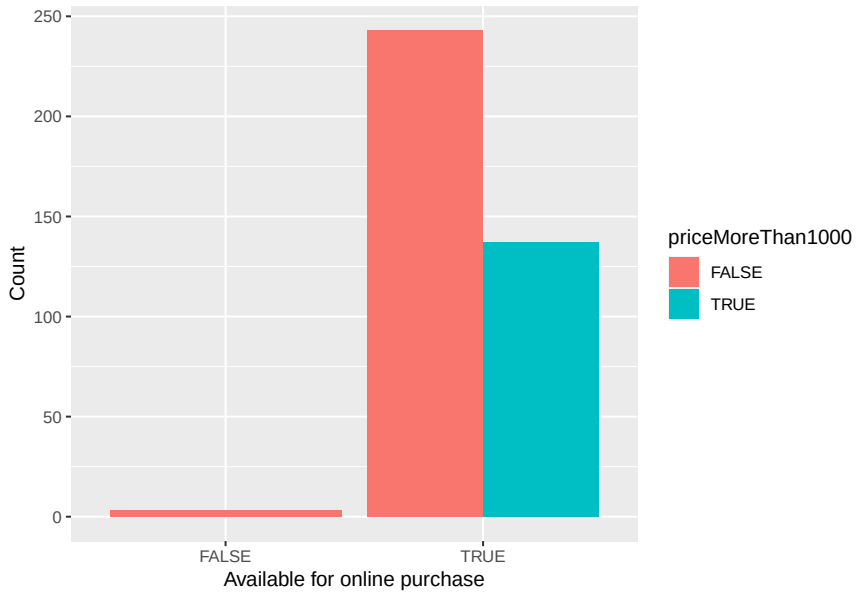
### 2.2.1 Sellable_online

First, we focus on the relationship between the `sellable_online` binary variable and `priceMoreThan1000`. The table of proportions for each `price category` within each `sellable_online` category computed below shows that all the products that are not available for online purchase cost 1000 Saudi Riyals or less, while products available online present a 36/64 split with inexpensive products still dominating.

However, from the barplot of counts along with our findings from the summary statistics it is shown that the amount of products not available for online purchase are just three in our data set. That's an extremely small amount of data as 99% or 380 of the products are indeed available for online purchase. In other words, we do not have a sufficient sample of products that are *unavailable* online in order to analyse the sample distribution of the price category for them. Therefore, under these data, we consider the `sellable_online` variable not significant for the analysis as all but 3 products are available online. It is important to point out that this issue was also present prior to cleaning the data as the original data set contained only 8 observations with `sellable_online` equal to `FALSE` out of 500 products in total.

Table 9: Percentages of price category within each level of sellable_online.

| | <=1000 | >1000 |
|---|---|---|
| not online | 1.0000000 | 0.0000000 |
| online | 0.6394737 | 0.3605263 |



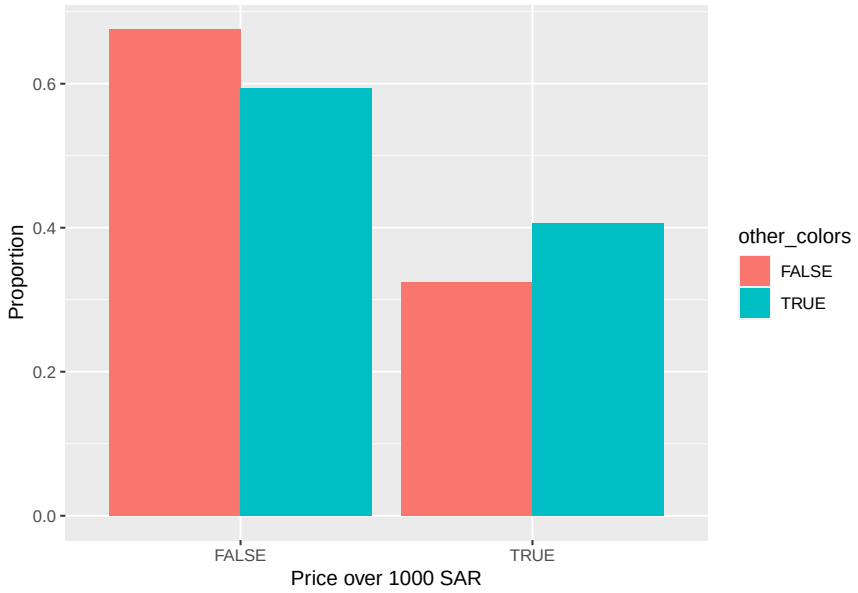

### 2.2.2 Other_colors

The barplot of the price category with respect to the `other_colors` binary variable and the table of the corresponding proportions indicate that the proportion of products available in multiple colours priced over

1000 SAR is larger than that of the products in one colour. However, these proportions are very similar. Furthermore, the proportion of single coloured items priced 1000 SAR or below is greater than those priced over 1000 SAR. We observe the same pattern for the category of multiple colours as well.

Table 10: Percentages of price category within each level of other_colors.

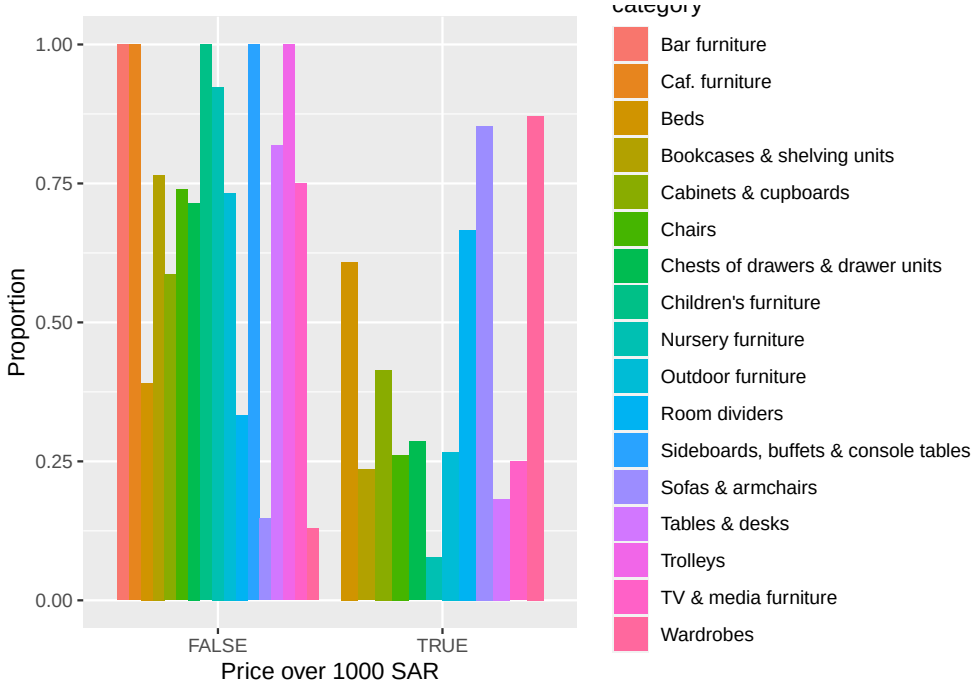
| | <=1000 | >1000 |
|---|---|---|
| not in other colours | 0.6754386 | 0.3245614 |
| in other colours | 0.5935484 | 0.4064516 |




### 2.2.3 Furniture category

The barplot of the price category by furniture category shows that for the most of the furniture categories the proportion of items priced at 1000 SAR or lower is larger than those priced over 1000 SAR. Especially, for the categories of "Bar furniture", "Caf.. furniture", "Children's furniture", "Sideboards, buffets & console tables" and "Trolleys", we observe that all of their products are priced at most 1000 SAR.

Finally, the only categories with larger percentages for items priced over 1000 SAR are "Beds", "Room dividers", "Sofas & armchairs" and "Wardrobes".




### 2.2.4 Height, width, depth

Before considering the price as a binary variable, it would be useful to explore its relationship to the other continuous variables.

One suitable summary statistic to assess the linearity of the relationships between continuous variables is the correlation coefficient. The correlation table below suggests a strong positive linear relationship between `price` and `width`, while there is a moderately strong positive linear relationship between price and depth. Additionally, the correlation between price and height is equal to 0.225, suggesting a very weak positive linear relationship.

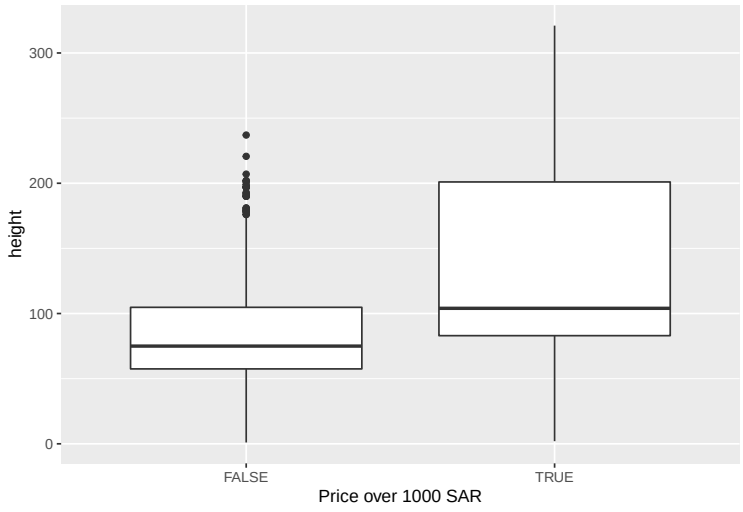
Finally, we observe that there is weak to moderate positive correlation between width and depth or height and that there is almost no correlation between height and depth.

Table 11: Correlation coefficients between price, width, height and depth.

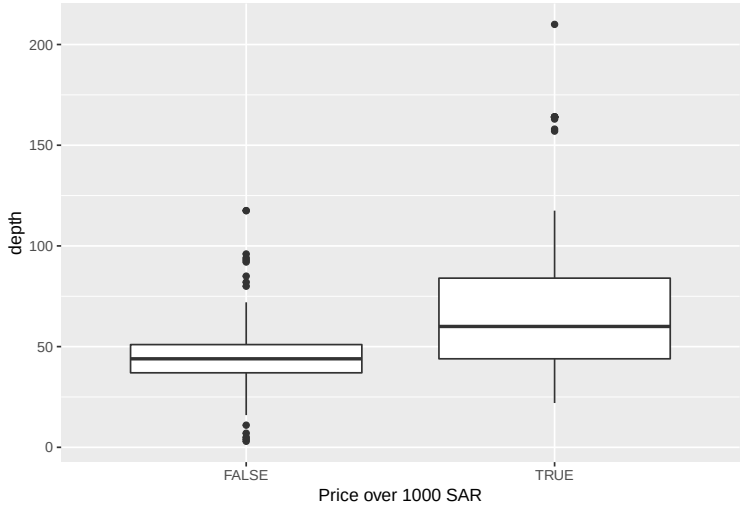
| | price | depth | height | width |
|---|---|---|---|---|
| price | 1.000 | 0.636 | 0.225 | 0.754 |
| depth | 0.636 | 1.000 | -0.042 | 0.478 |
| height | 0.225 | -0.042 | 1.000 | 0.406 |
| width | 0.754 | 0.478 | 0.406 | 1.000 |

We are now considering price as a binary variable.

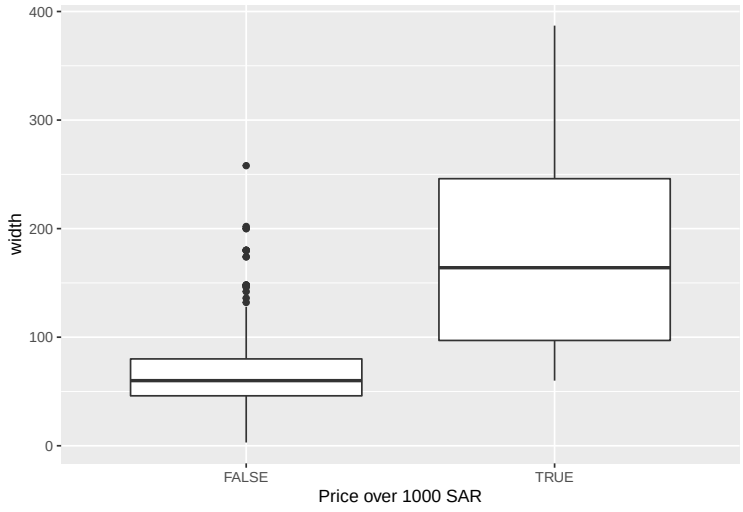
The boxplot of the sample distribution of height by the price category shows that the median value of height for items priced over 1000 SAR is greater than that of less expensive items, suggesting that the items priced over 1000 SAR have larger height, in general. Furthermore, the boxplot reports a lot of high outliers for the distribution of less expensive products, while the variance of height for expensive products is significantly larger.




The boxplot of depth by price category presents similar results. Items that cost more than 1000 SAR have in general more depth, while the variance of depth for those items appears to be significantly larger than that of the lower price items. Additionally, we observe several low and high outliers for the distribution of `priceMoreThan1000 = FALSE`, while there are some high outliers for the distribution of `priceMoreThan1000 = TRUE`.




Finally, we observe the largest separation in medians for the variable of width. Again, items priced over 1000 SAR are, in general, wider than those priced at 1000 or less. However, in the case of depth, the IQR boxes do not overlap at all as opposed to the previous 2 continuous variables. We observe several high outliers for the case of items which cost 1000 SAR or less, while the variance for items above 1000 SAR is considerably larger.




# 3 Formal Data Analysis

In this part, we will try to build the model, and make variable selection and model comparison. So as to establish a final model, and study the correlation between the variables and whether the price is higher than 1000.

## 3.1 Variable Selection

At the beginning, we will add all variables except "sellable_online" to the model. The model will be as follows:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_0 \cdot \text{depth} + \beta_1 \cdot \text{height} + \beta_2 \cdot \text{width} + \beta_3 \cdot \mathbb{I}_{other\_colors}(x)$$

where $p$ is the binary response variable which indicate the probability that the price is greater than 1000. $\mathbb{I}_{other\_colors}$ is an indicator variable such that

$$\mathbb{I}_{other\_colors}(x) = \begin{cases} 1 & , \ if\ x = \text{"TRUE"} \\ 0 & , \ \text{otherwise} \end{cases}$$

Table 12 shows the Confidence interval of all parameters.

Table 12: The Confidence Intervals of all parameters

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -7.5127761 | -4.8554418 |
| depth | 0.0213036 | 0.0567501 |
| height | 0.0033865 | 0.0143002 |
| width | 0.0186174 | 0.0319098 |
| other_colorsTRUE | -1.2366818 | 0.0853546 |

It can be seen that except for the variable "other_colors", all confidence intervals do not include 0. Therefore, "other_colors" cannot be considered significant. The high p-value can double-check the result.

Table 13: coefficients of model1

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.1010786 | 0.6757599 | -9.028471 | 0.0000000 |
| depth | 0.0381788 | 0.0090092 | 4.237770 | 0.0000226 |
| height | 0.0087735 | 0.0027747 | 3.161934 | 0.0015672 |
| width | 0.0248761 | 0.0033825 | 7.354447 | 0.0000000 |
| other_colorsTRUE | -0.5589168 | 0.3358748 | -1.664063 | 0.0960999 |

Therefore we decided to remove the variable "other_colors" and choose some of the remaining variables to buld the model. They are depth, height and width of each item in our dataset. In order to determine whether there are still variables that need to be removed, we compare AIC and BIC parameters to determine.

Table 14: Model comparison values for different models.

| Model | AIC | BIC |
|---|---|---|
| Keep all veriables | 278.64 | 294.43 |
| Remove width | 361.37 | 373.22 |
| Remove height | 361.37 | 373.22 |
| Remove depth | 296.99 | 308.84 |

It can be seen that AIC and BIC have the smallest values when we do not remove any variables. So our final model is the GLM model of whether the price is more than 1000 and the length, width and height of the products. In terms of expressions:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_0 \cdot \text{depth} + \beta_1 \cdot \text{height} + \beta_2 \cdot \text{width}$$

The prediction results for all parameters are as follows:

Table 15: coefficients of model1

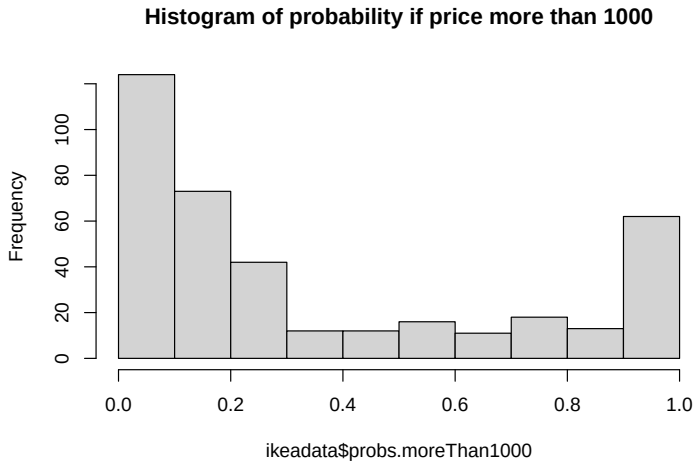
| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.1418797 | 0.6738822 | -9.114174 | 0.0000000 |
| depth | 0.0370486 | 0.0090129 | 4.110623 | 0.0000395 |
| height | 0.0086899 | 0.0027364 | 3.175653 | 0.0014950 |
| width | 0.0239085 | 0.0032704 | 7.310574 | 0.0000000 |

Our glm model estimates that the Intercept is -6.142. All of the intercept and slopes have low p-value which indicate the significance. As shown in the Statistical results, when we hold other variables in the model constant:

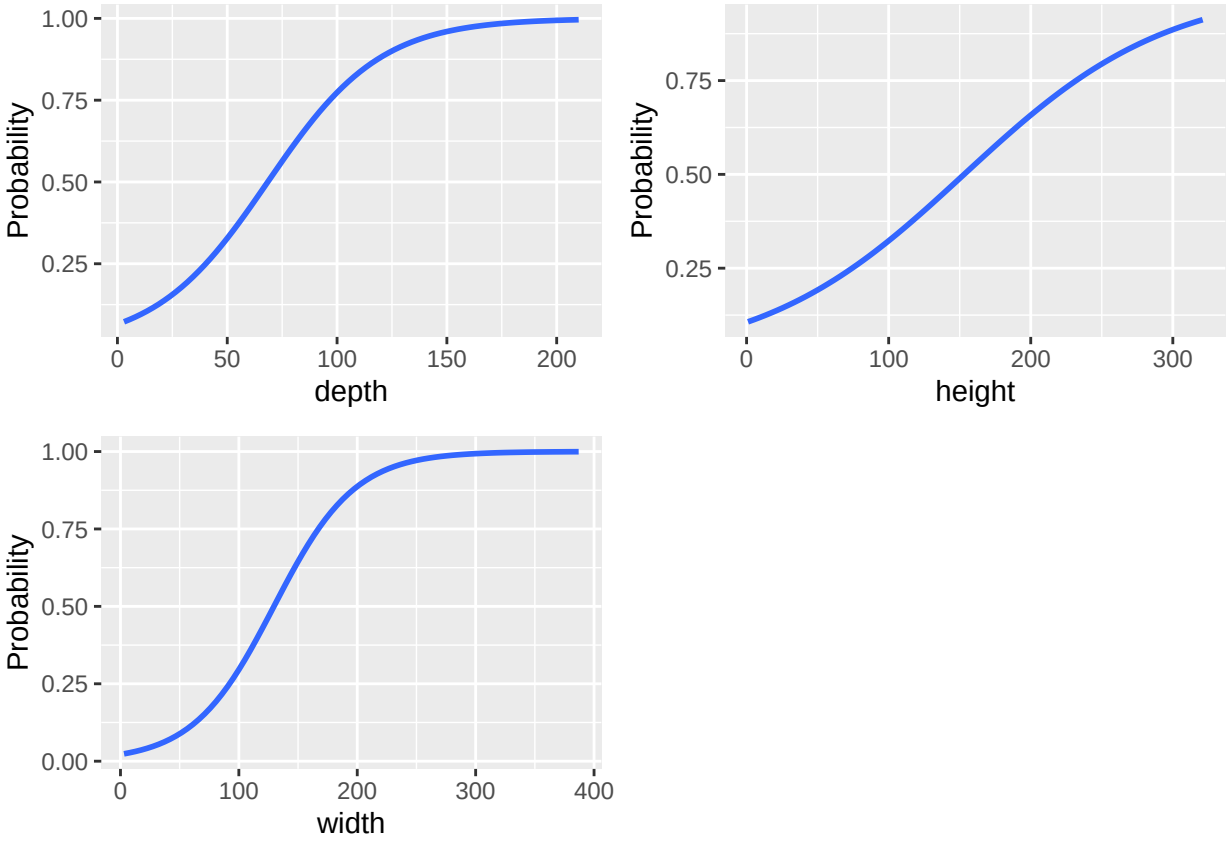
- the log-odds of the price of an item being over 1000 SAR increase by 0.037 for every one unit increase in `depth`
- the log-odds of the price of an item being over 1000 SAR increase by 0.009 for every one unit increase in `height`
- the log-odds of the price of an item being over 1000 SAR increase by 0.024 for every one unit increase in `width`

## 3.2 Show In Plot

We calculated the probability that the predicted price would exceed 1000. If it is greater than 0.5, it will be judged that the price exceeds 1000, and if it is less than 0.5, it will be judged that the price is less than 1000.




Then we draw the relationship between the remaining three variables and the probability. It can be seen that depth and width have a very obvious impact on the response variable, while the curve of height is relatively smooth.



# 4 Concluction

The results show that the price of furniture sold by IKEA is related to the size of furniture. The greater the depth, height and width of furniture, the greater the possibility of selling for more than 1000 Saudi Riyals. Moreover, the depth of furniture is the most influential attribute, and its increase will have the most impact on the increase of the possibility that the price exceeds 1000.

# 5 Future work

- Study about the data which contain 2 or more missing value.
- Is there any other algorithm as link function
- Is it possible to integrate category and reduce the number of categories so that it can be used as a variable in the model?