Group 23 Analysis

Group 23

Introduction

This project aims to determine which features of items sold by IKEA Saudi Arabia determine whether the items cost more or less than 1000 Saudi Riyals. This will be done by carrying out a binary logistic regression on a dataset containing information about items available from the retailer. The regression parameters will then give us information on which variables affect the probability of the price being above the threshold. Among those variables that do affect the probability, the relative size of the regression parameters will also allow us to see which variables affect the probability more than others.

Initial Data Cleaning and Exploration

Having inspected the .csv file, we can see that there are some cleaning steps to be done before the data is usable for modelling. First, let us look at the first rows of the dataset:

Table 1: The features of the first five items in the dataset.

item_id	category	price	sellable_online	other_colors	depth	height	width
90291698	Bookcases & shelving units	175	TRUE	No	NA	64	60
80280515	Chairs	225	TRUE	No	46	76	54
39287398	Bookcases & shelving units	340	TRUE	Yes	30	202	40
89135944	Sofas & armchairs	1995	TRUE	Yes	99	83	198
29248345	Bookcases & shelving units	906	TRUE	No	50	226	134

The first column, `item_id`, gives a numerical label for each item. This is unlikely to be related to the price of the item, so we should drop it from the dataset.

Category

The `category` column is currently presented as a column of strings, but there are a lot of repeated values, as shown in table 2

This means the column should be converted to a factor and treated as a categorical variable in the model. Given that categorical variables create a separate parameter in a model for each unique category, leaving this column in its current form may create a very complex model. We will rely on our model selection procedure to show us which parameters can be dropped from them model.

Price

This column contains the price of the item in Saudi Riyals, and will be the basis for our target variable. Our aim is to estimate the importance of the other variables in predicting whether an item costs more than 1000

Table 2: The distinct categories in the dataset and the number of items in each category, sorted from largest category to smallest.

Category	Number of Itens
Tables & desks	89
Bookcases & shelving units	71
Sofas & armchairs	58
Chairs	57
Cabinets & cupboards	44
Wardrobes	32
Beds	31
Outdoor furniture	29
TV & media furniture	28
Children's furniture	16
Nursery furniture	13
Chests of drawers & drawer units	10
Bar furniture	8
Cafe furniture	5
Room dividers	3
Sideboards, buffets & console tables	3
Trolleys	3

Riyals, i.e. whether the item's entry in column is above or below 1000. The distribution of this column is shown in figure 1.

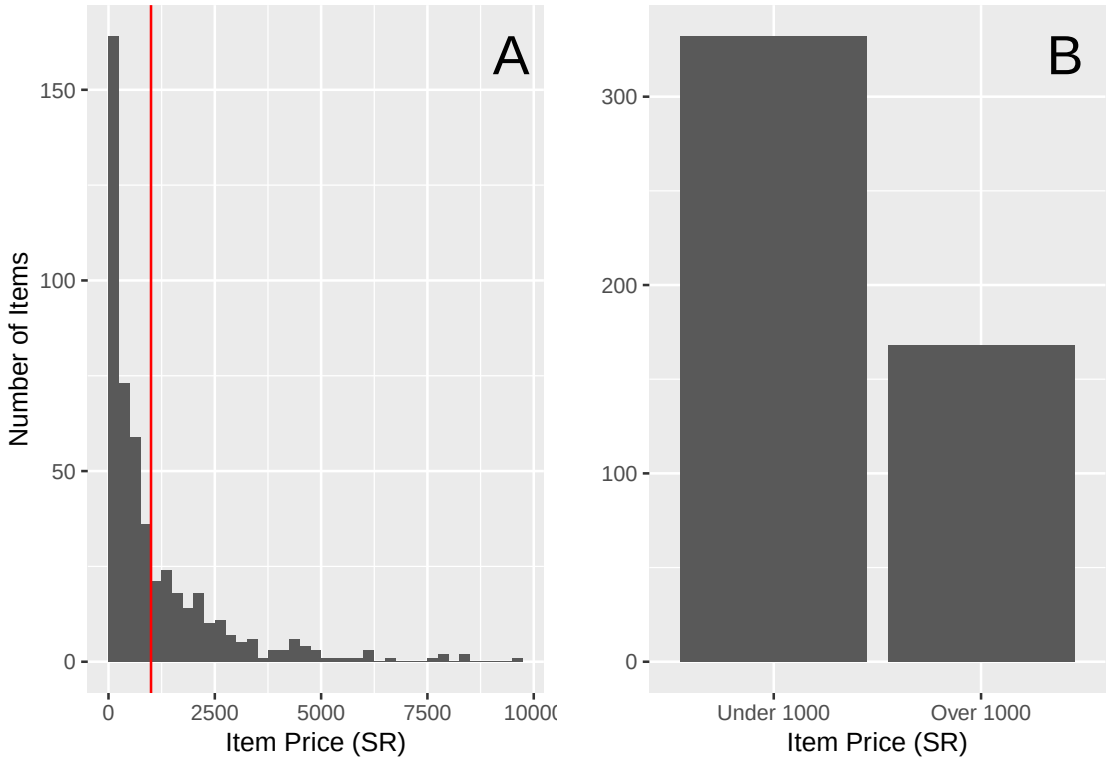

Figure 1: **A**: The distribution of prices, measured in Saudi Riyals (SR). Each bin is 250SR wide, and the red line marks 1000SR on the x-axis. **B**: The number of items with prices below 1000SR and above 1000SR.

From these graphs, we can see that the distribution of prices is quite skewed to the right. The first bar of graph A is the largest, so it is most common for items to be priced under 250SR. The bars generally get smaller as the price gets larger, but the distribution has a long tail out to around 9000SR. Graph B shows the number of items below and above 1000SR. It appears that there are around 320 items below 1000SR and around 160 items above 1000SR.

Sellable Online

The `sellable_online` column is a binary variable indicating whether the item can be purchased via the internet. Table 3 shows that this variable is very unbalanced: almost all of the items are available online. This may limit the usefulness of this variable when predicting the price category, but this will become more clear once the model has been fitted.

Table 3: The number of items available or unavailable online.

	Number of Items
Sellable Online	495
Not Sellable Online	5

Other Colors

The `other_colors` column is another binary variable, taking the value "yes" when the item is available in other colours and "no" when it is not. Table 4 shows how many items fit into each group. This column is more balanced than `sellable_online`, with about 40% of items available in another colour and 60% unavailable.

Table 4: The number of items available or unavailable in other colours.

	Number of Items
Not Available in Other Colours	304
Available in Other Colours	196

Depth, Height & Width

These three variables describe the physical dimensions of each item, measured in centimetres. As can be seen from the first row of table 1, these variables can contain missing values, so we will have to address this before using these variables for modelling. First we can look at the distribution of each variable in figures 2, 3 and 4.

Depth

From figure 2, it appears that the most common values are between 20 and 60cm, with the peak being somewhere between 30 and 50cm. The frequency of depths drops off quickly above 60cm, apart from a spike in the 90-100cm bin. There are a few items with depths of greater than 150cm, but most of the distribution occurs below 100cm.

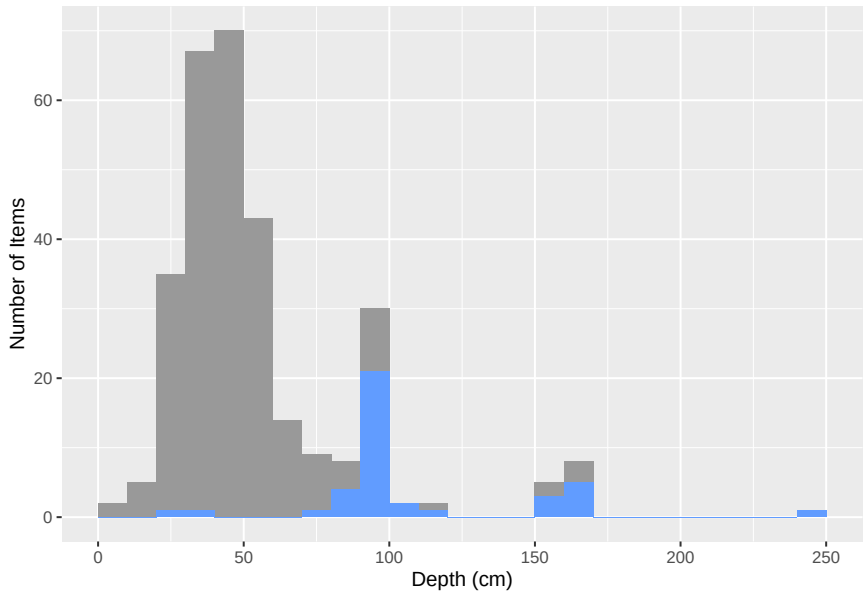

Figure 2: The distribution of depth measurements in cm. Each bin covers a 10cm range. The blue bars indicate the number of sofas and armchairs in each bin. The grey part of the bar shows the number of all other items.

Table 5: The proportion of items with depth between 90 and 100cm that are sofas, compared to items outside this depth range.

	Number of Items	Sofa Proportion
Depth Less Than 90cm or Greater Than 100cm	271	0.07
Depth Between 90 & 100cm	30	0.70

The unusually large number of items with depth between 90cm and 100cm is explained by a large number of sofas having depths in this range, as can be seen from 2. More precisely, table 5 shows that 70% of the items in this depth range are sofas or armchairs, compared to 7% in the rest of the distribution. Items like these may tend to have depths in this range to accommodate cushions, as well as a sitting person.

Table 6: The proportion of items with depth between 90 and 100cm that are sofas, compared to items outside this depth range.

	Number of Items	Sofa Proportion
Depth Less Than 150cm	287	0.11
Depth Greater Than 150cm	14	0.64

It can also be seen in figure 2 that the outlying values above 150cm are mostly composed of sofas and armchairs. Table 6 shows the exact proportions of items above and below 150cm in depth that are sofas and armchairs. From these numbers, we can see that items of this category are over represented at depth values over 150cm.

Height

Figure 3 shows that height is distributed more widely than depth, with some amount of the distribution present from 0 to 250cm. and There are typically 10-30 items in each of the bins between 30 and 110cm, apart from a strong peak between 70 and 90cm. Outside that range, there are typically between 1 and 10 items per bin throughout the variable range.

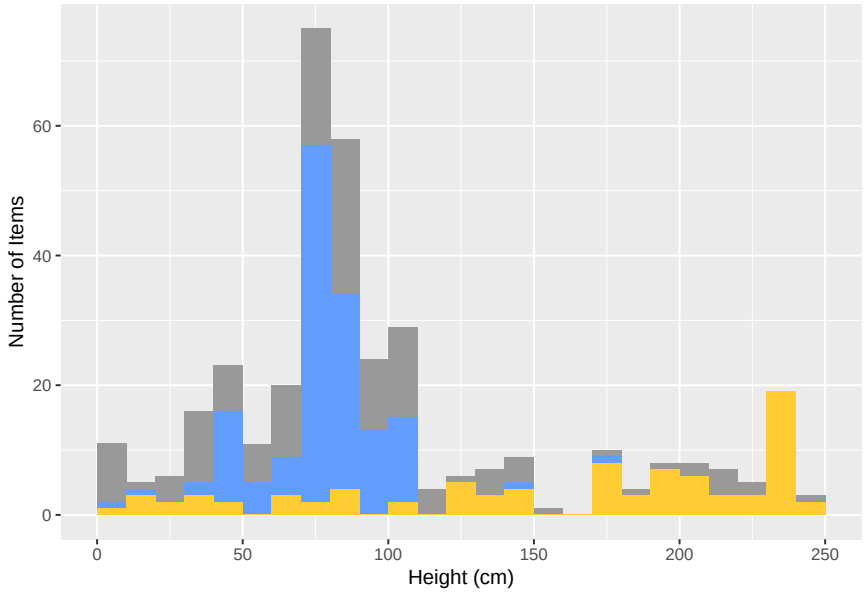

Figure 3: The distribution of height measurements in cm. Each bin covers a 10cm range. The blue component of each bar shows the number of items in each bin that are sofas, armchairs, tables, desks or chairs. The yellow part of each bar shows the number of wardrobes, bookcases or shelving units in each bin. The grey part of the bar shows the number of all other items.

Table 7: The proportion of items with height between 70 and 90cm that are sofas, armchair, chairs, tables or desks compared to items outside this height range.

Height (cm)	Number of Items	Chair Proportion	Sofa Proportion	Table Proportion
< 70 or > 90	228	0.06	0.10	0.08
70 - 90	141	0.15	0.16	0.31

From figure 3, we can see that a significant proportion of the items between 70 and 90cm in height are in the "Chairs", "Sofa & armchairs" or "Tables & desks" categories, whereas a smaller proportion of items outside this height range appear to fall into these categories. Table 7 shows the proportion of items in the peak of the height distribution that fall into the aforementioned categories, compared to items outside this height range. Within this height range, 15% of items are chairs (compared to 6% elsewhere), 16% are sofas or armchairs (compared to 10% elsewhere) and 31% are tables or desks (compared to 8% elsewhere). This pattern makes sense, as tables need to be of roughly a certain height in order to be useful to most people. Items for sitting on will likely be a similar height to tables, as we typically sit when using a table.

Table 8: The proportion of items over 165cm in height that are wardrobes, bookcases or other items, compared to the proportion for items below 165cm in height.

	Number of Items	Wardrobe Proportion	Bookcase Proportion
Height Less than 165cm	305	0.02	0.09
Height Greater Than 165cm	64	0.38	0.42

Figure 3 also shows that the upper part of the height range is dominated by items in the "Wardrobes" and "Bookcases & shelving units" categories. From table 8, wardrobes and bookcases form a significant proportion of the items over 165cm in height, accounting for 80% of such items in the dataset. They are

much less common at lower heights, which is to be expected given that these tend to be used for storing large objects (in the case of wardrobes) or centralising the storage of many small objects (in the case of bookcases).

Width

The distribution of width measurements is shown in figure 4. This distribution is somewhat similar to the distribution for depth in figure 2. There is a large peak roughly between 60 and 80cm, with a long tail to the right. The tail is heavier here than for depth, with items appearing all the way out to 360cm.

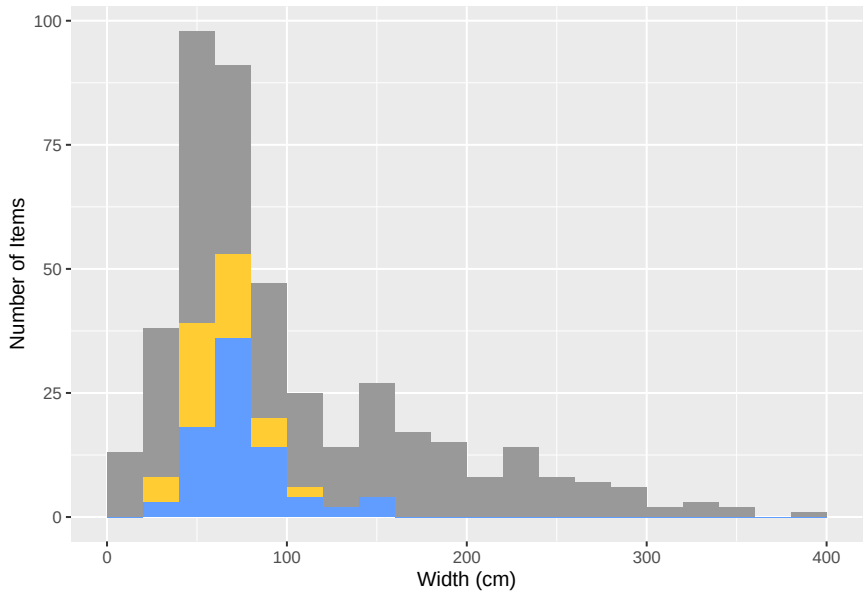

Figure 4: The distribution of width measurements in cm. Each bin covers a 20cm range. The yellow bars represent the number of chairs in each bin and the blue bars represent the number of tables or desks in each bin. The grey bars represent the number of all other items.

The distribution of width (shown in figure 4) appears to be more regular than that of depth or height. however, items are not uniformly represented throughout the distribution, as items from the "Chairs" and "Tables and desks" categories are present close to the peak of the distribution but absent in the tails.

Table 9: The proportion of items between 40 and 80cm in width that are chairs or tables and desks compared to these proportions for items outside this width range.

Width (cm)	Number of Items	Chair Proportion	Table Proportion
< 40 or > 80	232	0.05	0.11
40 - 80	204	0.20	0.27

From table 9, it can be seen that chairs and tables form a larger proportion of the items in the peak of the width distribution (between 40 and 80cm) than outside this range.

These distributions may have more structure hidden in them, as they include a range of items of many different categories. It is reasonable to expect that the dimensions of items in different categories would be distributed differently. For instance, we would expect wardrobes to typically be taller than chairs. We can examine the distributions of these variables for some of the larger categories listed in table 2.

Dimensions of Different Categories

The distributions of depth, height and width for each category are shown in figures 5, 6 and 7, respectively. The top plots show the 8 largest categories and the bottom plots show the rest.

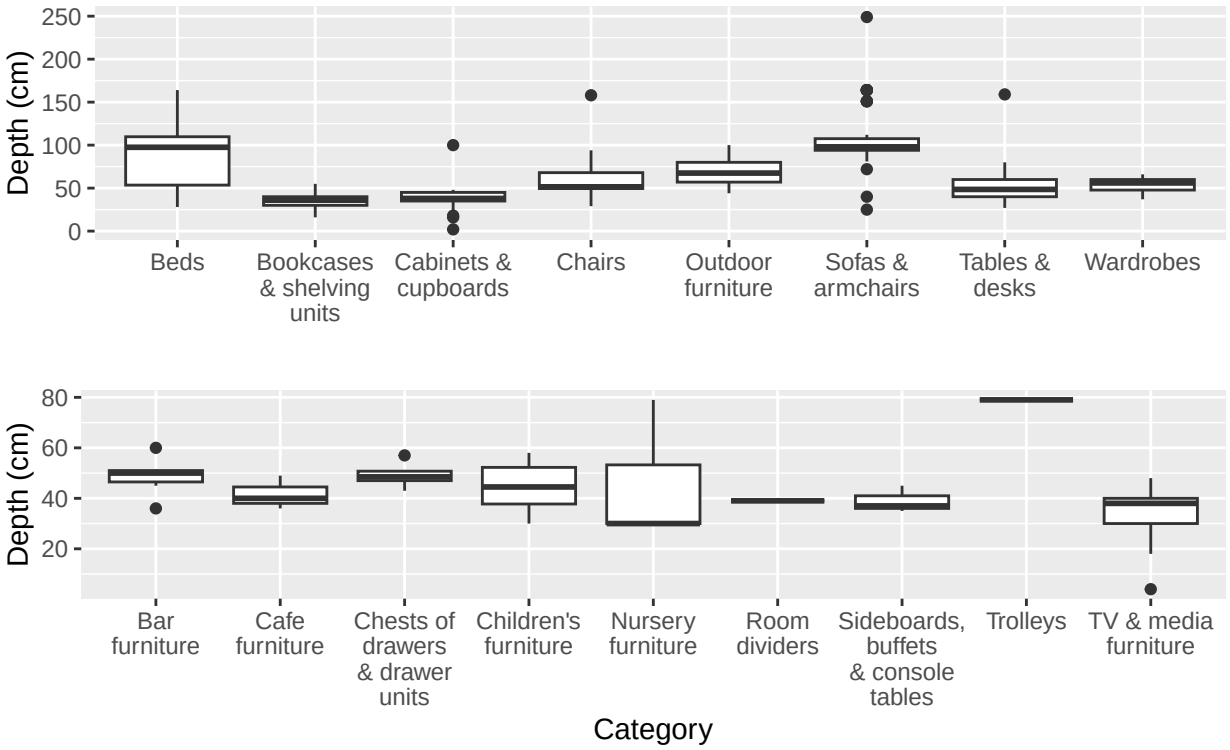

Figure 5: Boxplots of the depth (in cm) of items in each category.

From the upper plot in figure 5 we can see that the median depths for the more common categories are often around 50cm. The clear exceptions to this are the "Beds", "Outdoor Furniture" and "Sofas & Armchairs" categories. For the less common categories, the median depth is also typically close to 50cm, except for "Nursery Furniture" and "Trolleys". These categories have only a small number of items in each, so the distributions are likely to be noisier than for more numerous categories. There is also some difference in the range of these distributions, particularly for "Beds". This may be due to double and single beds both being included in this category.

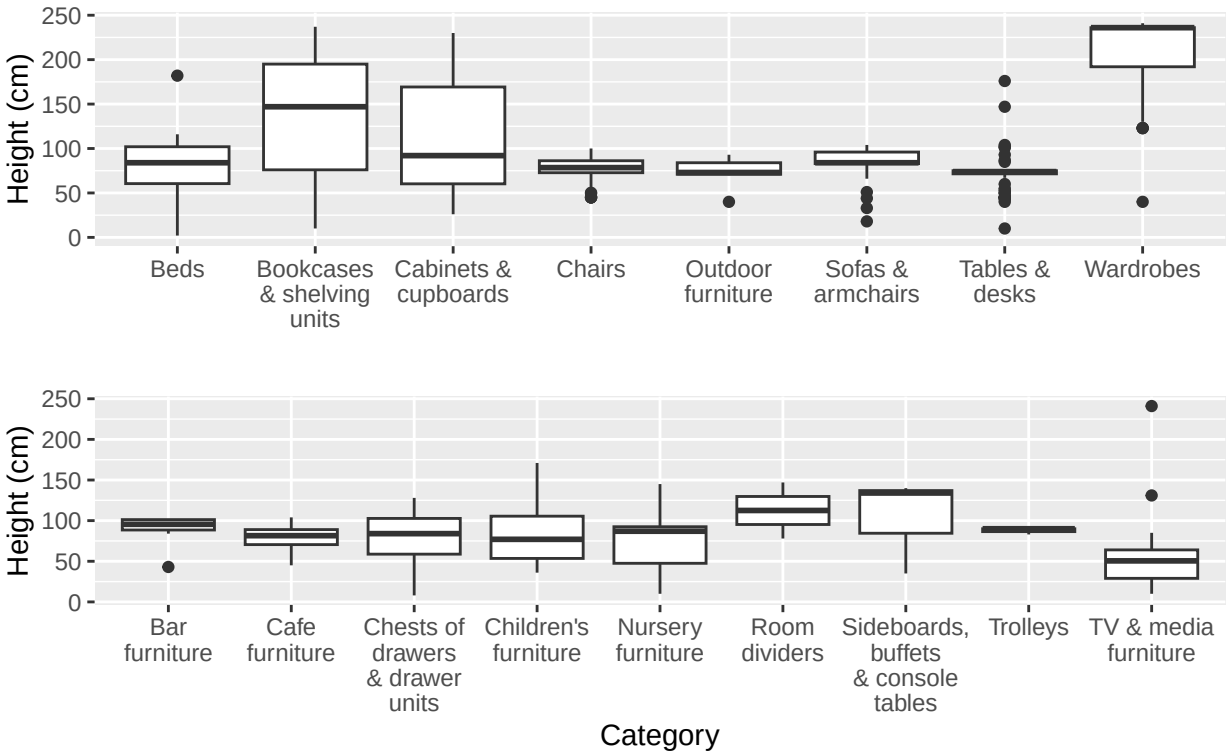

Figure 6: Boxplots of the height (in cm) of items in each category.

Figure 6 shows the boxplots of height for each category. In the upper plot, we can see that the median height for several of the categories is approximately 75cm, with the exception of "Bookcases & shelving units", "Cabinets & cupboards" & "Wardrobes". Interestingly, the median height of wardrobes is very close to the maximum height for this category, indicating that there are many items with the same or similar height. In the lower plot, the median heights are typically somewhat similar, between 75 and 100cm with the exception of "Room dividers", "Sideboards, buffets & console tables" (with medians over 100cm) and "TV & media furniture" with median height around 50cm.

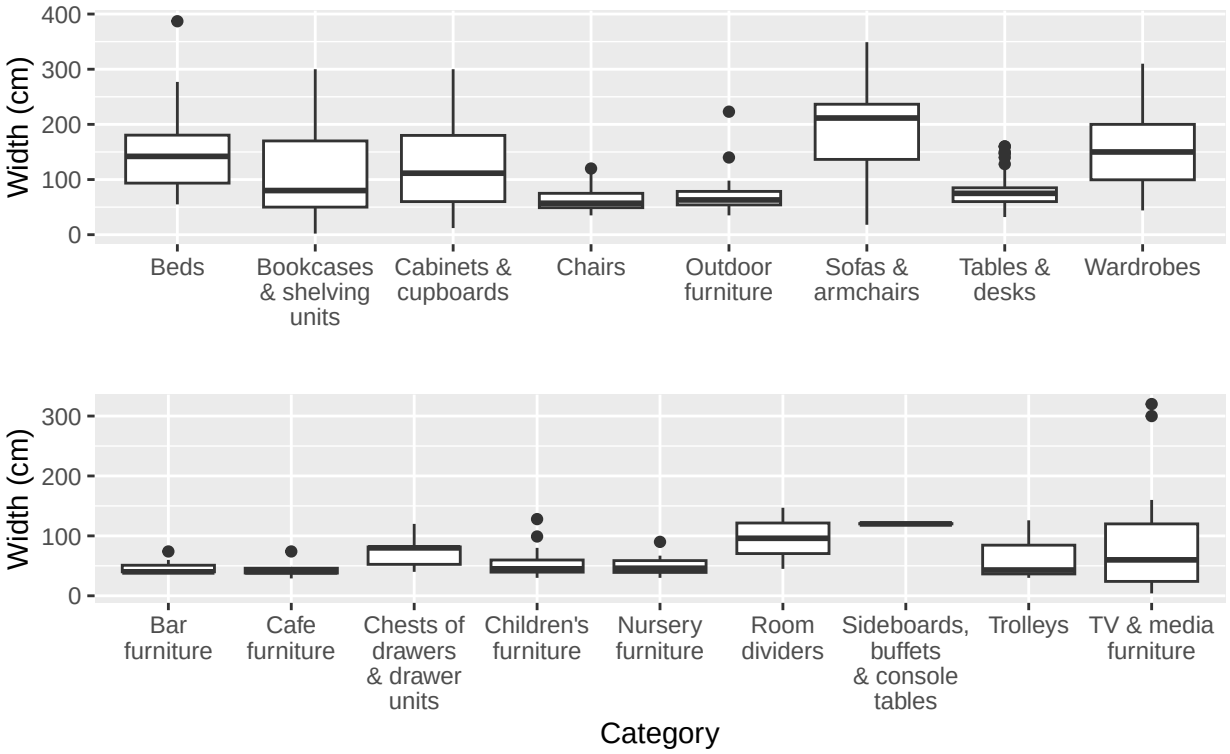

Figure 7: Boxplots of the width (in cm) of items in each category.

Lastly, figure 7 shows the boxplots of width for each category. In the upper plot, the medians range from approximately 50cm ("Chairs" and "Outdoor furniture") to much larger values: "Beds" and "Wardrobes" have medians of around 150cm and "Sofas & armchairs" has a median of over 200cm. The ranges of width values for the categories in this plot are often quite wide, for instance "Sofas & armchairs" ranges from below 50cm to over 300cm. For the lower plot, the medians appear a little more consistent, typically around 50cm, although "Chests of drawers & drawer units", "Room dividers" and "Sideboards, buffets & console tables" have noticeably larger medians.

Missing Data

Table 10: The number of missing values in each column and the proportion of each column that is not missing.

Variable Name	Number of Missing Values	Proportion of Non-Missing Values
category	0	1.000
other_colors	0	1.000
sellable_online	0	1.000
item_id	0	1.000
price	0	1.000
depth	199	0.602
height	131	0.738
width	64	0.872

Some of the items in the dataset are missing values for some variables, as can be seen from the first row of table 1. Table 10 shows the number of missing values in each column, along with the proportion of items that are not missing a value in that column. The only variables with missing information are the ones describing the dimensions of the item: depth, height and width. These missing values must be handled somehow if the data is to be used for modelling. The simplest way to treat this data would be to drop all rows which are missing one or more values. However, it can be seen from 10 that only 60.2% of rows are not missing their depth value. This means that dropping rows with missing values would remove at least 39.8% of the dataset. In fact, taking all three dimension variables into account, this cleaning strategy would remove 49.6% of the rows in the dataset. This seems like too much data to lose, so we must impute the missing values somehow. The distributions of these variables have some slightly irregular features and outliers, so mean imputation may be inappropriate. The median may be a more robust approach, although a global median for each variable may also be inappropriate as the distribution of each variable can differ strongly between item categories, as shown in figures 5, 6 and 7. Therefore, we can try cleaning these columns by assigning the median of the relevant item category in place of missing values.

Multiple Missing Values

In some cases, all 3 dimension variables are missing from an item. This occurs in 0.1% of rows. As this is a small proportion of the overall dataset and imputing all 3 dimensions of the item would likely lead to a poor approximation of the true values, we should drop these rows from the dataset we use for modelling.

Cleaning the Data

We are now in a position to prepare the dataset for modelling. The processing steps to be done are:

- Discard the `item_id` column
- Convert `category`, `sellable_online` and `other_color` to be factors
- Convert the `price` column into a binary target variable based on whether it is greater than 1000SR
- Drop any rows where all 3 of `depth`, `height` and `width` are missing
- Replace any remaining missing values of `depth`, `height` and `width` with the median value for their category.

Having applied these steps, we can see the features of the cleaned data. Comparing tables 1 and 11 shows that the transformations appear to have worked: the price is now a binary variable matching what we'd expect from the value in table 1 and the missing value for `depth` in the first row has been replaced, while the categorical variables appear unchanged apart from being converted into factors in the background.

Table 11: The features of the first five items in the dataset after the cleaning steps have been applied.

category	price	sellable_online	other_colors	depth	height	width
Bookcases & shelving units	Below 1000SR	TRUE	No	36	64	60
Chairs	Below 1000SR	TRUE	No	46	76	54
Bookcases & shelving units	Below 1000SR	TRUE	Yes	30	202	40
Sofas & armchairs	Above 1000SR	TRUE	Yes	99	83	198
Bookcases & shelving units	Below 1000SR	TRUE	No	50	226	134

Modelling

We can now begin considering our model. The first step here is to examine the relationship between each explanatory variable and the response variable.

Relationship Between Variables and Response

The data is now ready to be used for modelling. Before fitting the model, we can examine the relationship between the covariates and the response variable.

Category

First, we can see the proportion of items in each category that are above or below 1000SR, shown in figure 8. There is a lot of variation between the different categories, even between categories with many items. For instance, Sofas & armchairs has the greatest proportion of items above 1000SR (around 75%), whereas for Chairs the proportion is roughly 25%. From this, we may expect that category will be a useful predictor of whether an item has price above 1000SR.

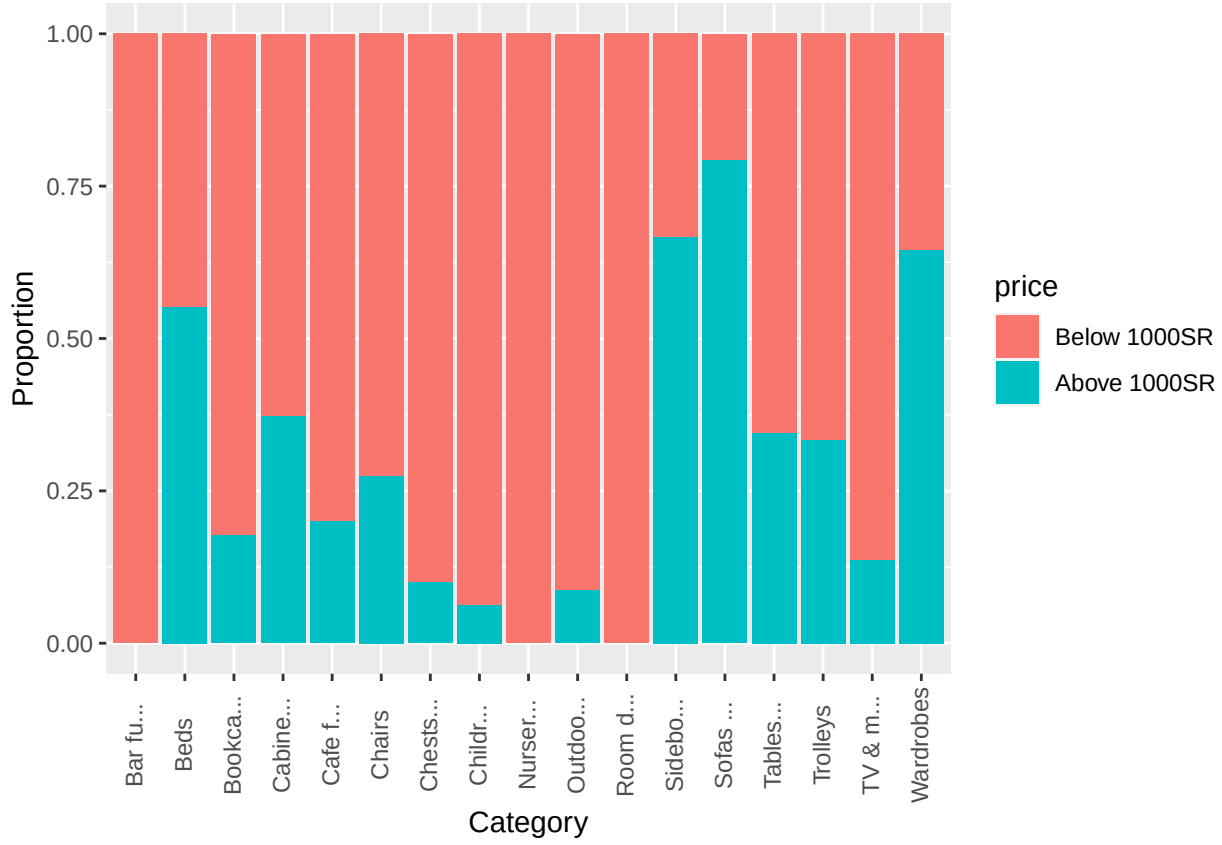

Figure 8: The proportion of items in each category with price below or above 1000SR.

Sellable Online

As figure 9 shows, all items which are not available online have price below 1000SR. This deviation from the population price distribution is not surprising, as there are only 5 items in this group (see table 3). As a result, it seems unlikely that this variable will be useful in predicting the response variable.

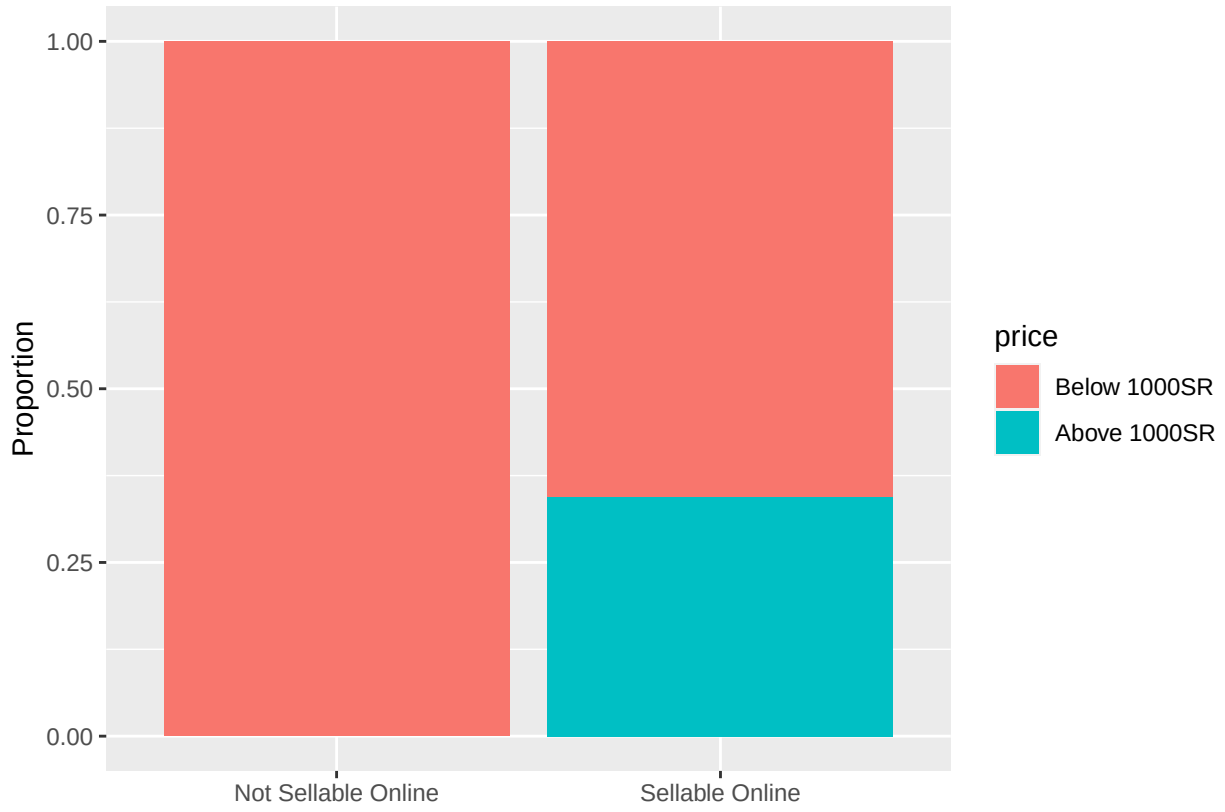

Figure 9: The proportion of items with price below or above 1000SR, grouped by whether they are sellable online or not.

Other Colors

Figure 10 shows the distribution of response variable values for items grouped by whether they are available in other colors or not. It appears that items available in multiple colors are more likely to have prices above 1000SR (about 45%) than those available in just one color (closer to 25%). Based on this, this variable may be a moderately strong predictor of whether the item's price is above the threshold.

Figure 10: The proportion of items with price below or above 1000SR, grouped by whether they are available in other colors or not.

Depth

For each value of the response variable, figure 11 shows a boxplot of the depth distribution. The upper quartile for items in the lower price bracket is just over 50cm, whereas the lower quartile of the more expensive items is just below 50cm and the upper quartile is near 100cm. This suggests that an item with a large depth value will have a price above 1000SR.

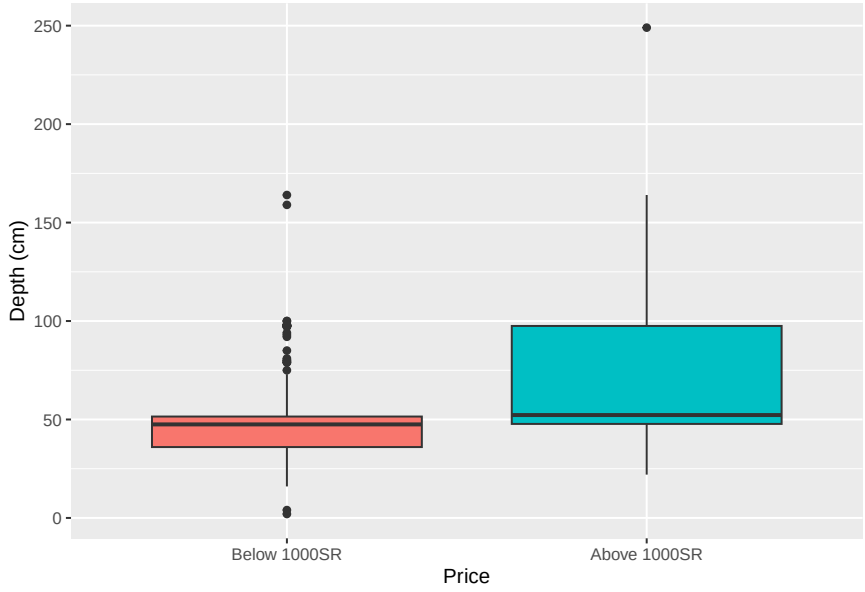

Figure 11: Boxplots of depth for items with price below or above 1000SR.

Height

Similarly, we can plot the height distributions for the two values of the response variable. This is done in figure 12. Again, a greater height appears to indicate a greater probability of an item being in the higher price range. However, the distinction between the groups is less clear here, with the upper quartile of the "Below" group entirely overlapping the lower quartile of the "Above" group and outliers from the "Below" group extending all the way to the top of the range of the "Above" group. This variable may be a useful indicator of the response variable, but perhaps to a lesser degree than depth.

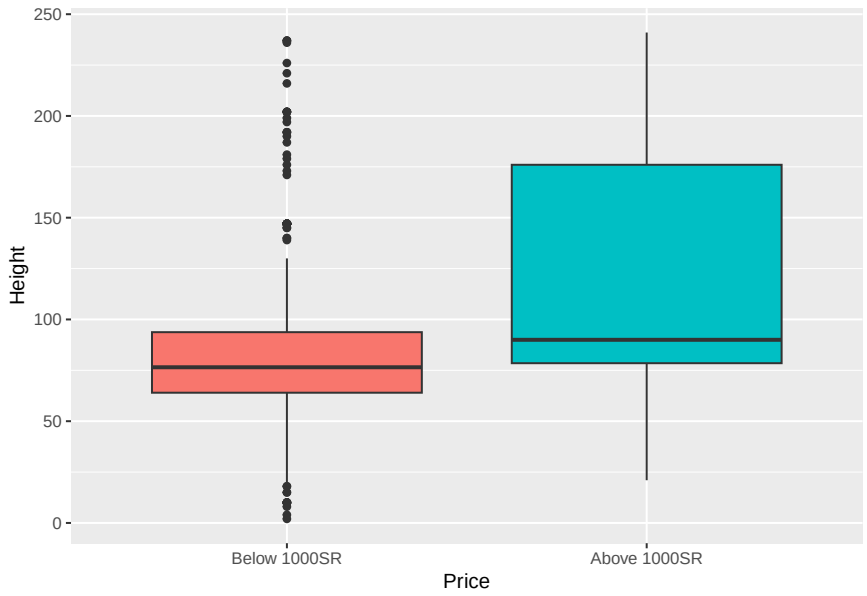

Figure 12: Boxplots of height for items with price below or above 1000SR.

Width

Lastly, we plot the distribution of width for the two values of the response variable in figure 13. This shows a similar pattern to figures 11 and 12: The lower priced items generally have smaller width values. The difference between the groups is more distinct than for height, but there are still outliers from the lower priced group extending through much of the higher priced group's range. This variable may be a fairly strong predictor of the response: perhaps stronger than height but weaker than depth.

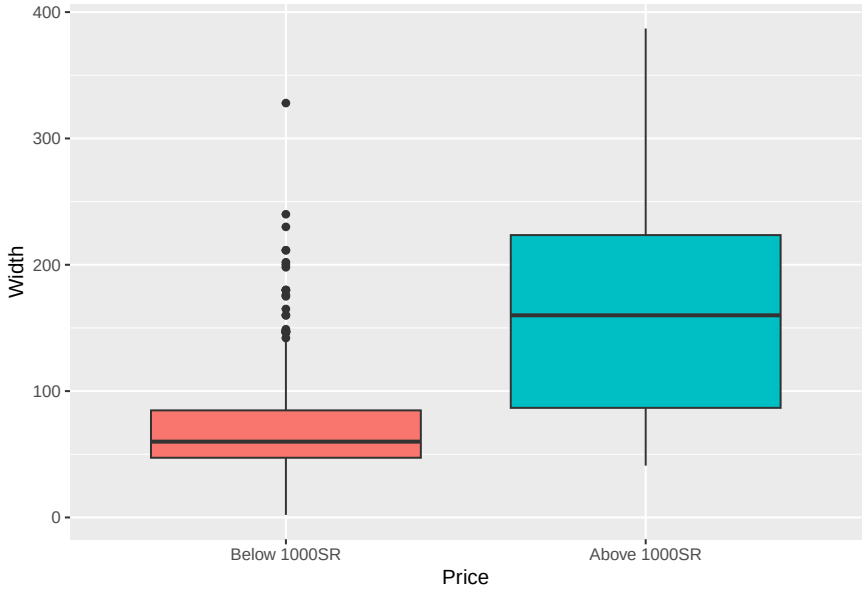

Figure 13: Boxplots of width for items with price below or above 1000SR.

Model Selection

At last, we can fit some models to investigate our question of interest: determining which variables influence whether an item will cost more than 1000SR. To do this we will fit a binary GLM with the price column as the target variable. The regression coefficients will then give the change in log-odds associated with a unit change in the corresponding covariate. Comparing the coefficients, then, will show us which variables are most important in determining whether an item costs more than 1000SR. The first model to fit will be a model incorporating all of the available covariates.

Full Model

This model is of the form:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j}^{16} \beta_j \mathbb{I}_j(x_i) + \beta_{17}\mathbb{I}_{sellable}(x_i) + \beta_{18}\mathbb{I}_{colours}(x_i) + \beta_{19}x_{i,depth} + \beta_{20}x_{i,height} + \beta_{21}x_{i,width}$$

where:

- p_i is the probability of the ith item having price above 1000SR
- β_0 is the intercept term
- $\beta_1, \beta_2, ..., \beta_{21}$ are the regression parameters
- $\mathbb{I}_1(), \mathbb{I}_2(), ..., \mathbb{I}_{16}()$ are indicator functions evaluating to 1 if the ith item is in the corresponding category. For the 17 categories, 16 indicators are required as the baseline category is represented by all indicator functions returning zero.
- $\mathbb{I}_{sellable}()$ is an indicator function returning 1 if the ith item is sellable online and 0 otherwise
- $\mathbb{I}_{colours}()$ is an indicator function returning 1 if the ith item is available in other colors and 0 otherwise
- $x_{i,depth}, x_{i,height}, x_{i,width}$ are the depth, width and height of the ith item in centimetres.

Table 12 shows the fitted parameter estimates for this model. We can see that the confidence intervals parameters associated with each of the different category values are all very wide and include zero. This

Table 12: The parameter estimates and associated confidence intervals from the first model.

	Parameter Estimate	95% CI Lower Bound	95% CI Upper Bound
(Intercept)	-36.452	-3881.188	3808.283
categoryBeds	15.072	-2630.028	2660.171
categoryBookcases & shelving units	10.890	-2634.210	2655.989
categoryCabinets & cupboards	14.683	-2630.416	2659.782
categoryCafe furniture	16.236	-2628.864	2661.336
categoryChairs	16.171	-2628.928	2661.270
categoryChests of drawers & drawer units	14.450	-2630.650	2659.549
categoryChildren's furniture	13.982	-2631.118	2659.082
categoryNursery furniture	-0.274	-3318.357	3317.809
categoryOutdoor furniture	14.020	-2631.079	2659.120
categoryRoom dividers	-3.268	-5224.789	5218.253
categorySideboards, buffets & console tables	16.111	-2628.990	2661.212
categorySofas & armchairs	15.588	-2629.512	2660.687
categoryTables & desks	16.375	-2628.724	2661.474
categoryTrolleys	15.781	-2629.319	2660.882
categoryTV & media furniture	14.232	-2630.868	2659.332
categoryWardrobes	12.341	-2632.759	2657.440
sellable_onlineTRUE	14.739	-2775.501	2804.979
other_colorsYes	0.007	-0.602	0.617
depth	0.010	-0.008	0.028
height	0.028	0.017	0.040
width	0.027	0.019	0.034

indicates that, when controlling for the other variables in the model, the category has no statistically significant effect on the propability of an item costing more than 1000SR. Given the differences seen in figure 8, it is surprising that none of the category parameters are significant, but it may be the case that the differences between categories are explained by other variables.

As category contributes a lot of complexity to the model and no significant parameters, we shall drop it and re-fit the model.

Model Without Category

The model fitted here contains all the same terms as the full model, except the terms involving $\beta_1, \beta_2, ..., \beta_{16}$. The parameter estimates are shown in table 13. The confidence intervals for sellable_online and other_colors, while smaller than they were in the first model, both include zero. Therefore these variables still do not have a significant effect on the log-odds of an item costing more than 1000SR. We can see that depth, width and height all have confidence intervals that do not include zero. We can now fit a model with just these three variables.

Model with Only Numeric Variables

The final model fitted here is one including only the depth, width and height covariates. These all had significant parameters in the previous model, so we expect them to still be significant now the other variables have been removed. From table 14, we can see that all parameters in this model (including the intercept) have confidence intervals that do not include zero, and are therefore significant at the $\alpha = 0.05$ level.

Table 13: The parameter estimates and associated confidence intervals from the second model.

	Parameter Estimate	95% CI Lower Bound	95% CI Upper Bound
(Intercept)	-17.727	-1244.705	1209.252
sellable_onlineTRUE	13.056	-1213.922	1240.034
other_colorsYes	0.009	-0.513	0.530
depth	0.022	0.011	0.033
height	0.007	0.002	0.012
width	0.019	0.014	0.024

Table 14: The parameter estimates and associated confidence intervals from the third model.

	Parameter Estimate	95% CI Lower Bound	95% CI Upper Bound
(Intercept)	-4.700	-5.593	-3.806
height	0.007	0.002	0.012
width	0.019	0.014	0.024
depth	0.022	0.012	0.033

Model AIC Comparison

The performance of the models can be compared by using the Akaike Information Criterion (AIC), a metric that rewards better fitting models but penalises models with many parameters. A lower value of AIC indicates a better model. Of the three models fitted here, the first (with all parameters) had the best AIC value, 359.28. The next best was the third model, with only depth, width and height as explanatory variables, with an AIC value of 395.63. The second model (excluding only the category variable) had the worst AIC value, 398.8. This is somewhat surprising, as all the extra parameters in the first model were not significant, and so might be expected to contribute little to the fit of the model while being strongly penalised by the AIC calculation. The third model outperforming the second model is more expected, as it is a smaller model that ignores some insignificant parameters. Despite the full model having the best AIC value, we decide not to use it as it does not help answer the research question; the category parameters are insignificant and may obscure the effect of other covariates.

Parameter Comparison

Comparing the parameters, we can see that depth has the largest estimated parameter, with a value of 0.022. This means that for every extra centimetre in depth, we expect the log-odds of the item costing over 1000SR to increase by 0.022. This is equivalent to multiplying the odds of this outcome by 1.023. The next largest parameter is for width, which gives the expected log-odds increase per centimetre of width as 0.019, or an odds multiplier of 1.019. The smallest parameter of the three is height, giving a log-odds increase of 0.007 (odds multiplier of 1.007) per centimetre.

The fact that these are the only variables that seem to have an effect on whether the price of the item is greater than 1000SR makes some sense, as the dimensions of the item determine roughly how much material went into making it and how costly it is to transport. The same amount of raw materials would plausibly cost the same regardless of whether you were building a bookcase or a bed. The relative sizes of the parameters is also somewhat expected, given the distributions seen in figures 11, 12 and 13, as the depth distribution was the most clearly separate between the price groups, followed by the width distribution and then the height distribution.

Conclusions

Fitting binary GLMs to the data and iteratively removing insignificant parameters indicates that the only variables influencing whether an item costs more than 1000 Saudi Riyals, are the depth, width and height of the item. Depth influenced the probability most, followed by width and then height. The category of item did not significantly affect the probability, despite different categories having quite different proportions of items in each price bracket. This may be due to the category correlating with the item size, but this modelling work could be extended to investigate this further. For instance, instead of including a parameter for each category in the model, the category variable could be converted into dummy variables and these variables could be removed one at a time during model selection, rather than all at once. This may lead to some informative categories being retained. As the results here indicate that the item size is the determining factor in whether its price is above 1000SR, a further extension to the work could be to multiply depth, width and height together to create a "volume" variable and fit a model using this covariate. This would indicate whether it is just overall "size" that determines price, and not the individual dimensions separately.