Group 23 Analysis

Group 23

2023-03-08

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 could try grouping some categories together (perhaps categories with similar median prices), or 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 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.

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

category	number
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

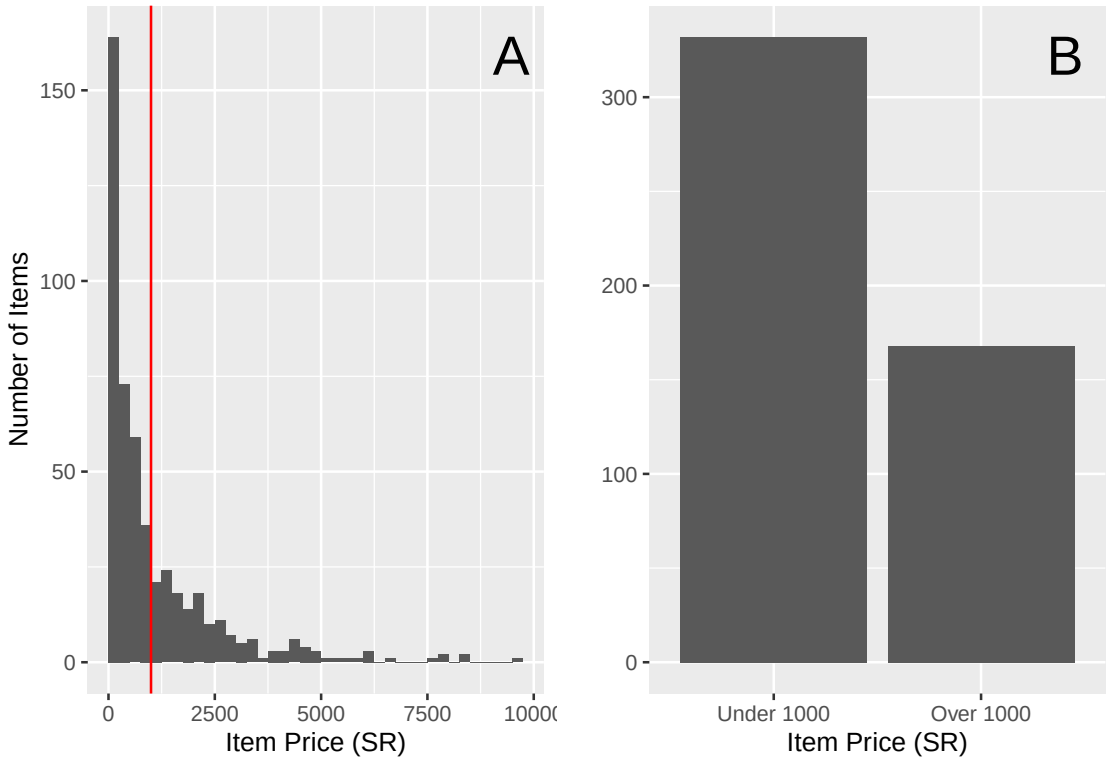

Figure 1: **A**: The distribution of prices, measured in Saudi Riyals (SR). Each bin is 250SR wide, and the red line marks 1000SR on the x-axis. **B**: The number of items with prices below 1000SR and above 1000SR.

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.

sellable_online	number_of_items
TRUE	495
FALSE	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.

other_colors	number_of_items
No	304
Yes	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 at the 90-100cm bin. There are a few items with depths of greater than 150cm, but most of the distribution occurs below 100cm.

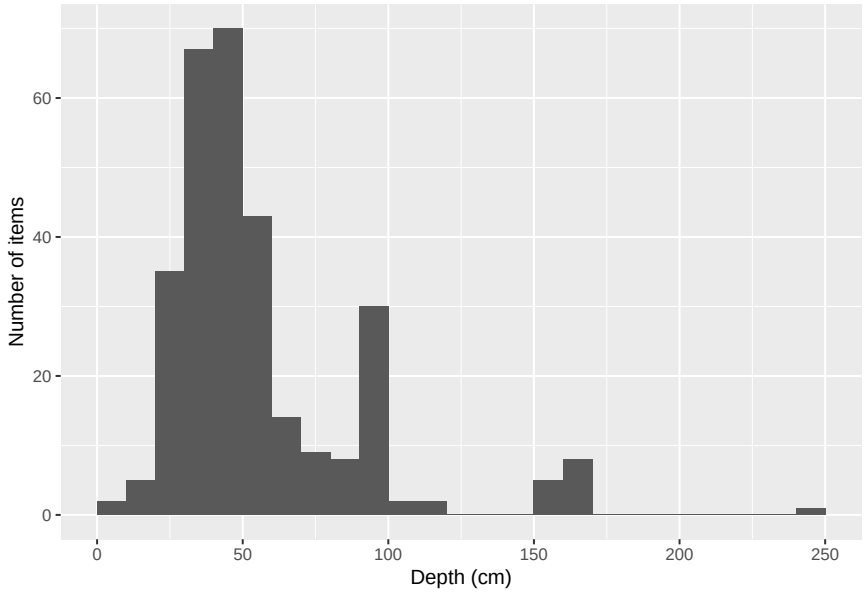

Figure 2: The distribution of depth measurements in cm. Each bin covers a 10cm range.

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

is_90_100	total	sofa_proportion
FALSE	271	0.07
TRUE	30	0.70

The spike in depth between 90cm and 100cm is explained by a large number of sofas having depths in this range, as can be seen from table 5

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

is_outlier	total	sofa_proportion
FALSE	287	0.11
TRUE	14	0.64

From table 6, it is evident that the large depth values are also largely because of sofas and armchairs.

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.

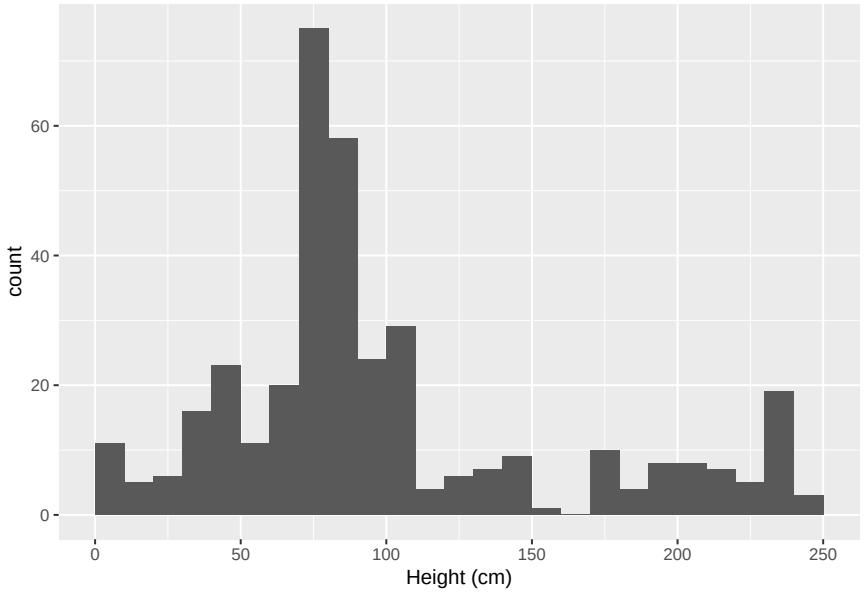

Figure 3: The distribution of height measurements in cm. Each bin covers a 10cm range.

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

is_peak	total	chair_prop	other_prop	sofa_prop	table_prop
FALSE	228	0.06	0.76	0.10	0.08
TRUE	141	0.15	0.38	0.16	0.31

Table 7 shows the proportion of items in the peak of the height distribution that are chairs, sofas or tables or none of these ("other" items), compared to items outside this height range. We can see that sofas, chairs and particularly tables are overrepresented in this height range compared to elsewhere. This 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.

is_upper	total	wardrobe_prop	other_prop	bookcase_prop
FALSE	305	0.02	0.89	0.09
TRUE	64	0.38	0.20	0.42

Evidently, 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.

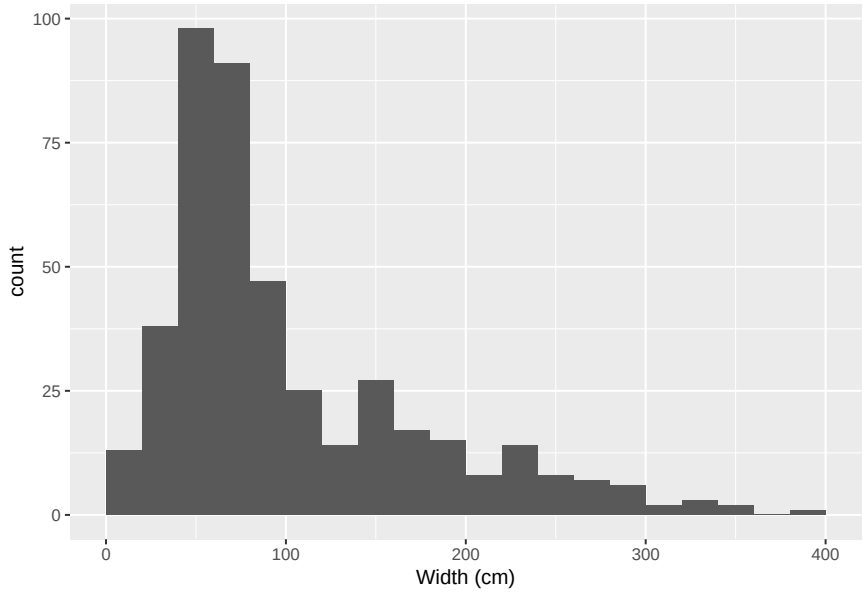

Figure 4: The distribution of width measurements in cm. Each bin covers a 20cm range.

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 listen in table 2.

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

is_peak	total	chair_prop	other_prop	table_prop
FALSE	232	0.05	0.84	0.11
TRUE	204	0.20	0.53	0.27

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

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.

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.

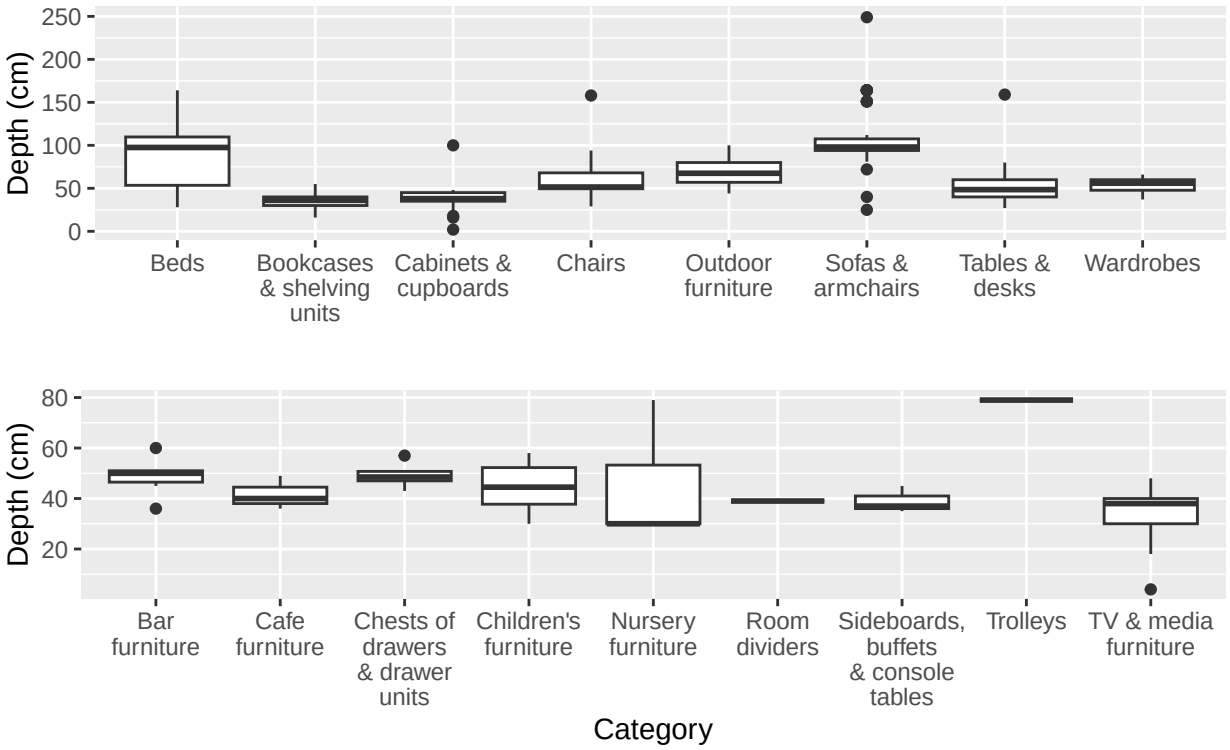

Figure 5: Boxplots of the depth (in cm) of items in each category.

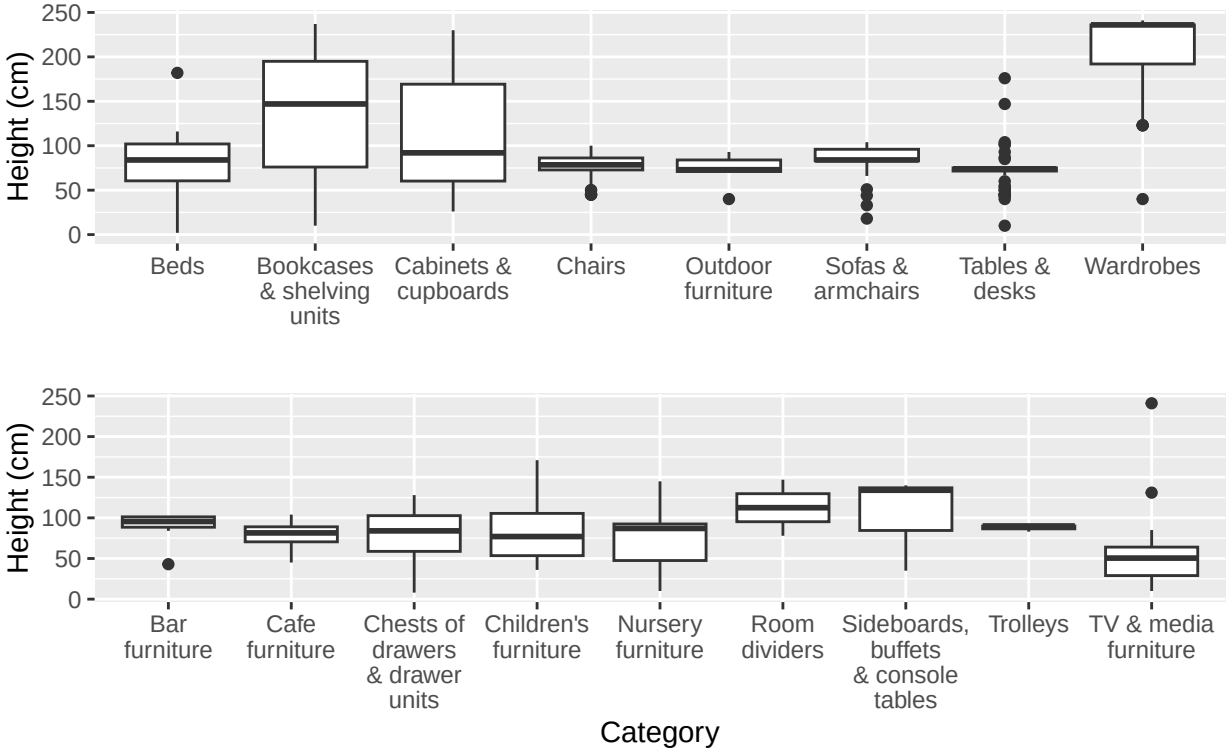

Figure 6: Boxplots of the height (in cm) of items in each category.

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.

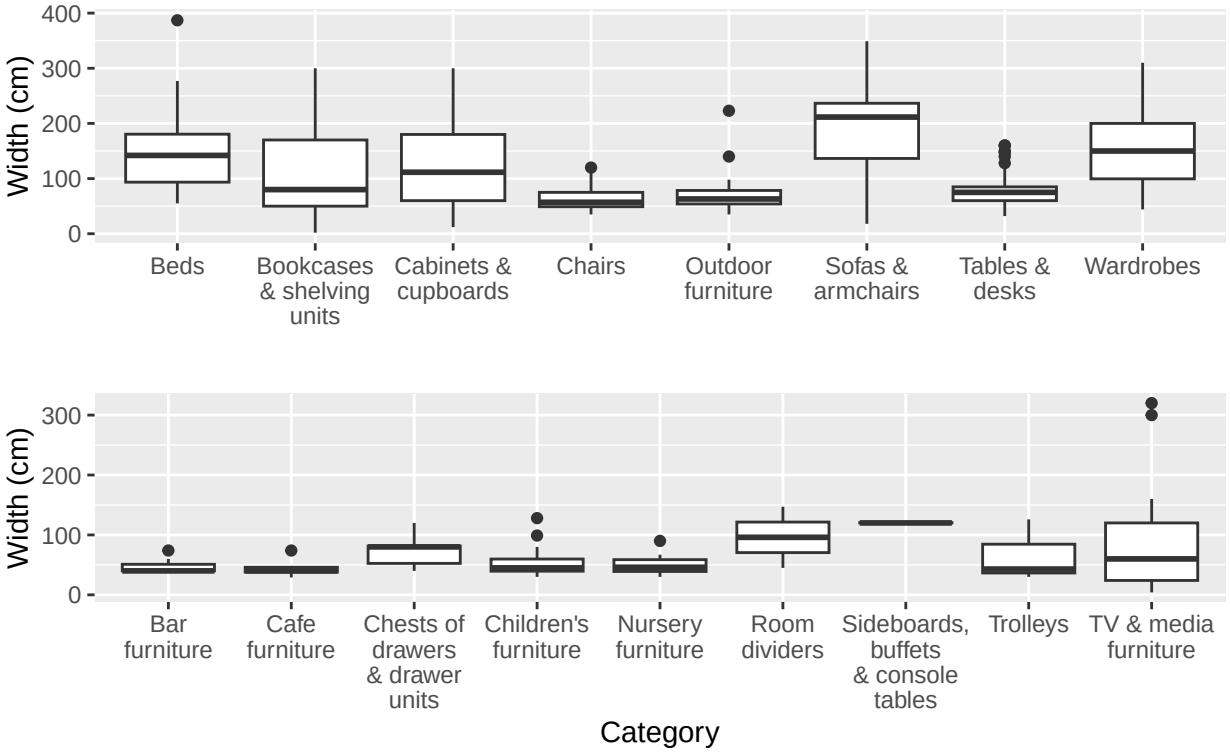

Figure 7: Boxplots of the width (in cm) of items in each category.

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, with some tanging 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

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 1 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

Table 10:

skim_variable	n_missing	complete_rate
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

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.

These steps are applied below:

```
cleaned_dataset <- raw_dataset %>%
  # filter out rows missing all 3 columns
  filter(!if_all(c("depth", "height", "width"), is.na)) %>%
  select(-item_id) %>% #drop the item id column
  mutate( # most of the changes are simple vectorised conversions
    category=as.factor(category),
    sellable_online=as.factor(sellable_online),
    other_colors=as.factor(other_colors),
    price=as.integer(if_else(price >= 1000, 1, 0))
    )

# replacing missing values will be more complicated
# there may be a cleaner way to do this
for (cat in category_counts$category){ #loop through each category in the dataset

  cat_filter <- cleaned_dataset$category == cat  # find all items of this category
```

```
  for (i in 5:7){ # for each dimension column (columns 5, 6 and 7)

    # find all items with NA in this column
    na_filter <- is.na(cleaned_dataset[, i])

    # replace the values matching both filters with the median for this category
    cleaned_dataset[cat_filter & na_filter, i] <- median(cleaned_dataset[cat_filter, i],
                                                         na.rm=TRUE)
  }
}

head(cleaned_dataset, n=5) %>%
  kable(caption = '\\label{tab:cleaned} The features of the first five items in the dataset after the cl
   kable_styling(font_size = 10, latex_options = "hold_position")
```

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	0	TRUE	No	36	64	60
Chairs	0	TRUE	No	46	76	54
Bookcases & shelving units	0	TRUE	Yes	30	202	40
Sofas & armchairs	1	TRUE	Yes	99	83	198
Bookcases & shelving units	0	TRUE	No	50	226	134

Finally 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.