# GROUP 23

# DATA EXPLORATION

# 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 |
| | | | ... | | | | |

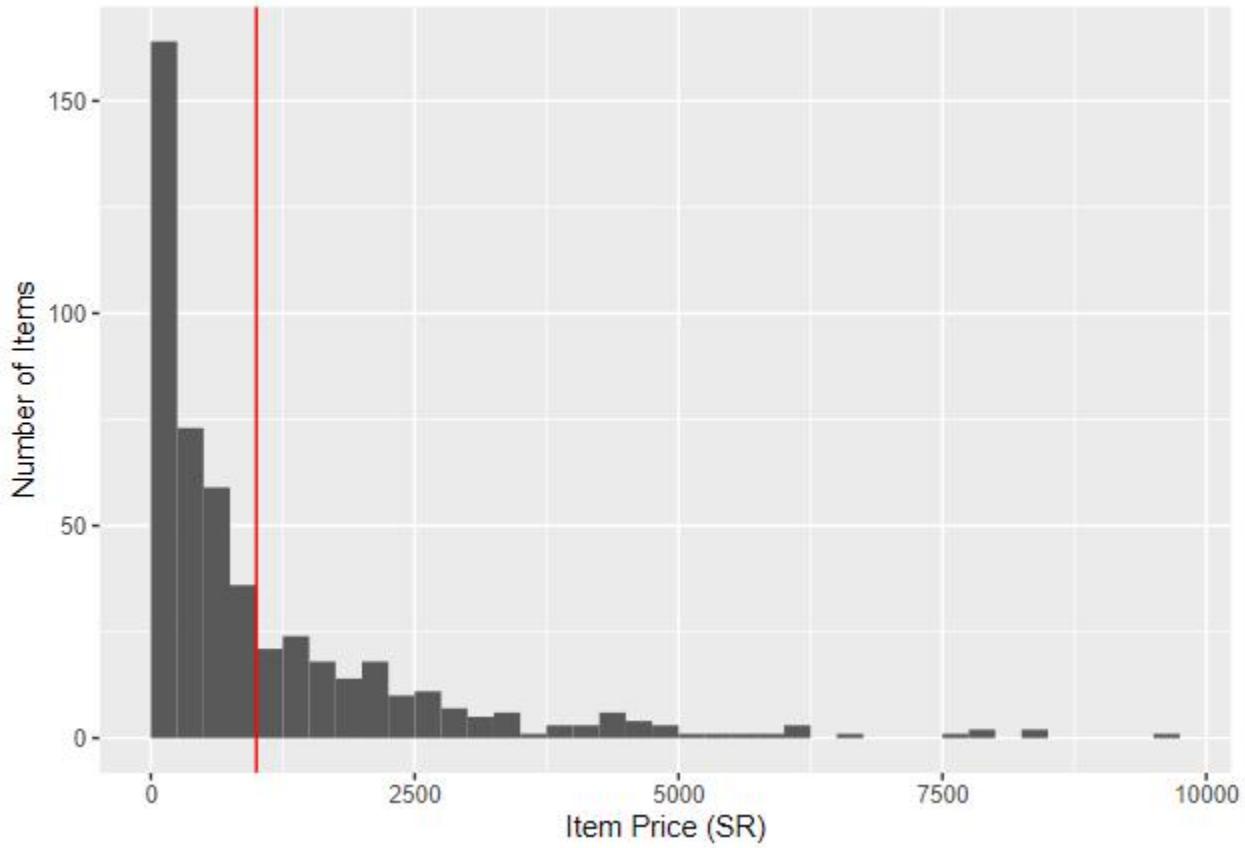
# CATEGORY

| 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 |

…

# PRICE

- Our target variable.
- Aim: Which properties of furniture influence whether they cost more than 1000 Saudi Riyals?



# SELLABLE ONLINE & OTHER COLORS

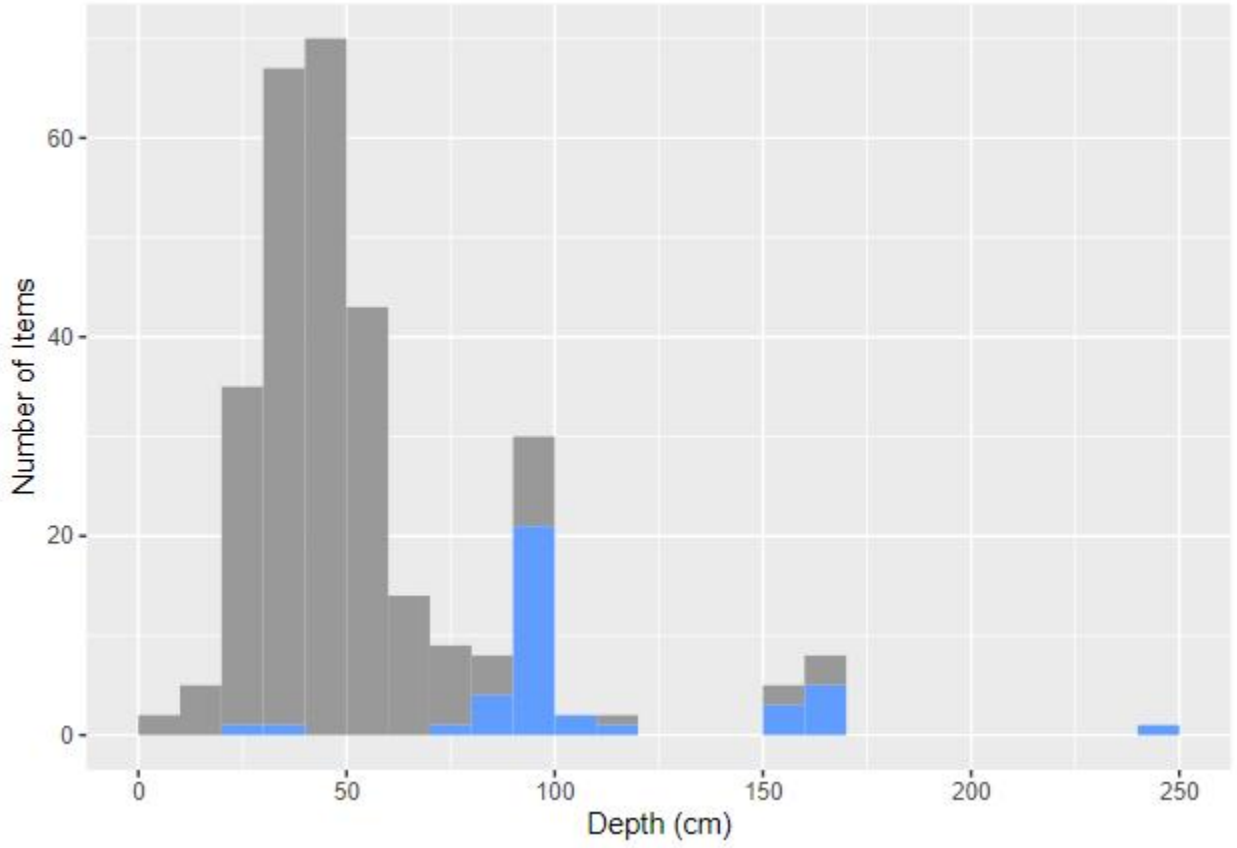
- Binary variable.
- Out of 500 items, 495 items are available online.
- The sellable online variable may not be significant due to the large difference in quantity.
- 304 items do not have other colors.
- The distribution of other colors is more balanced, it might give us more information.

# DEPTH & HEIGHT & WIDTH

- These three variables describe the physical dimensions of each item.
- Numeric variable, measured in centimetres.

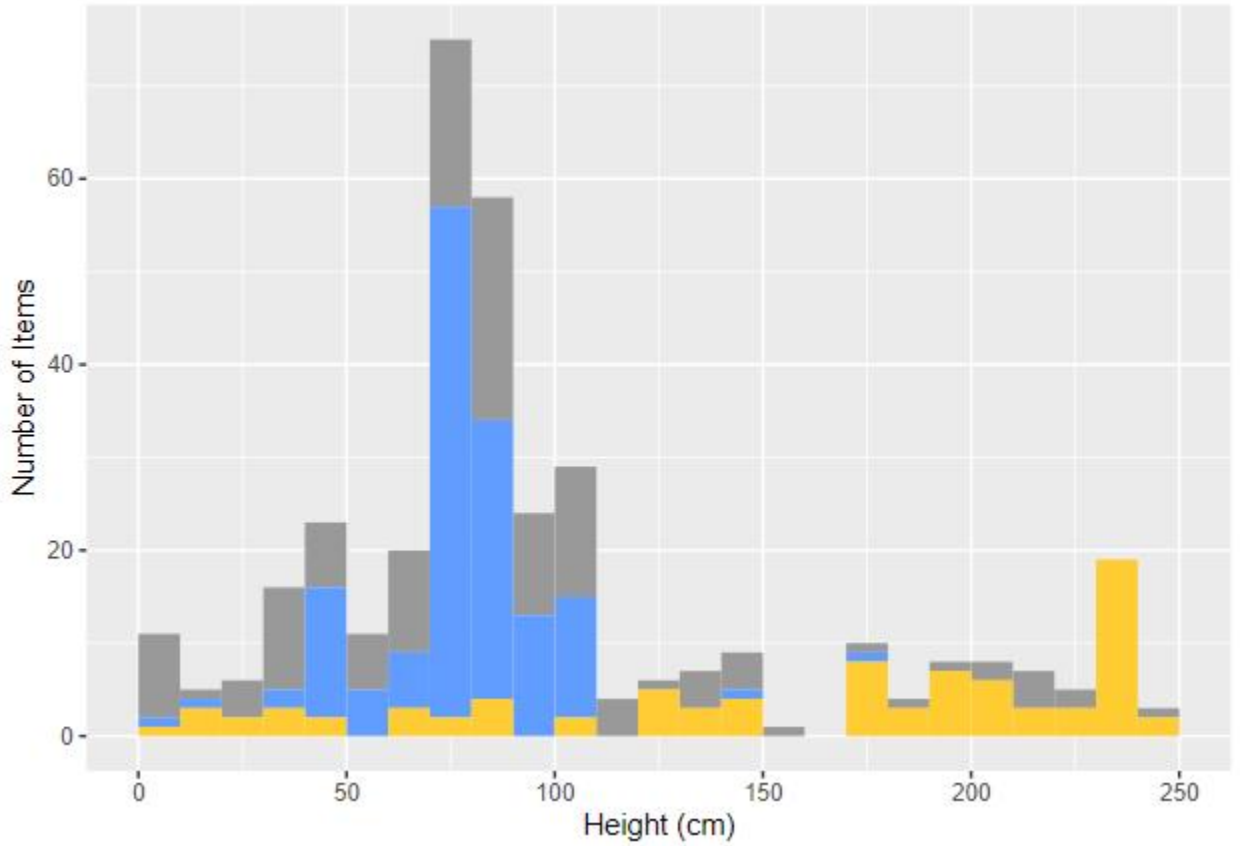
# DEPTH

- The most common values of depth are between 20cm and 60cm.
- The peak of depth is located around 30cm to 50cm.
- The frequency of depths drops off quickly above 60cm.
- The value of spike between 90cm and 100cm is 30, and the proportion of sofas (blue) in the spike is 70%.



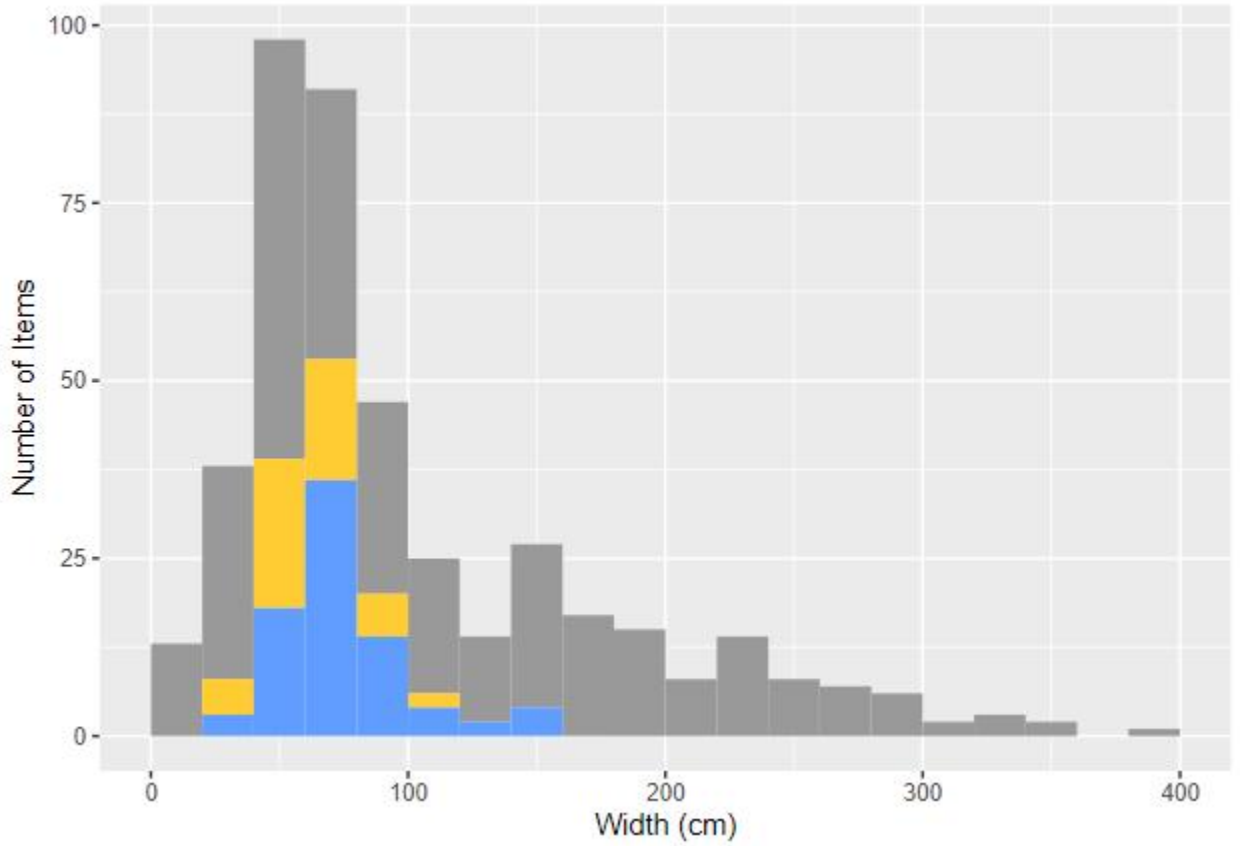
# HEIGHT

- The values of height are widely distributed from 0cm to 250cm.
- The peak is located between 70cm and 90cm.
- The items with the highest proportion of peaks are chairs, sofas or tables (blue).
- Wardrobes and bookcases form a significant proportion of the items over 165cm in height (yellow).



# WIDTH

- The width values are widely separated from 0cm to 400cm.
- The distribution has a long tail to the right.
- The peak is roughly located between 60cm and 80cm.
- Chairs (yellow) and tables (blue) form a high proportion of the items in the peak.



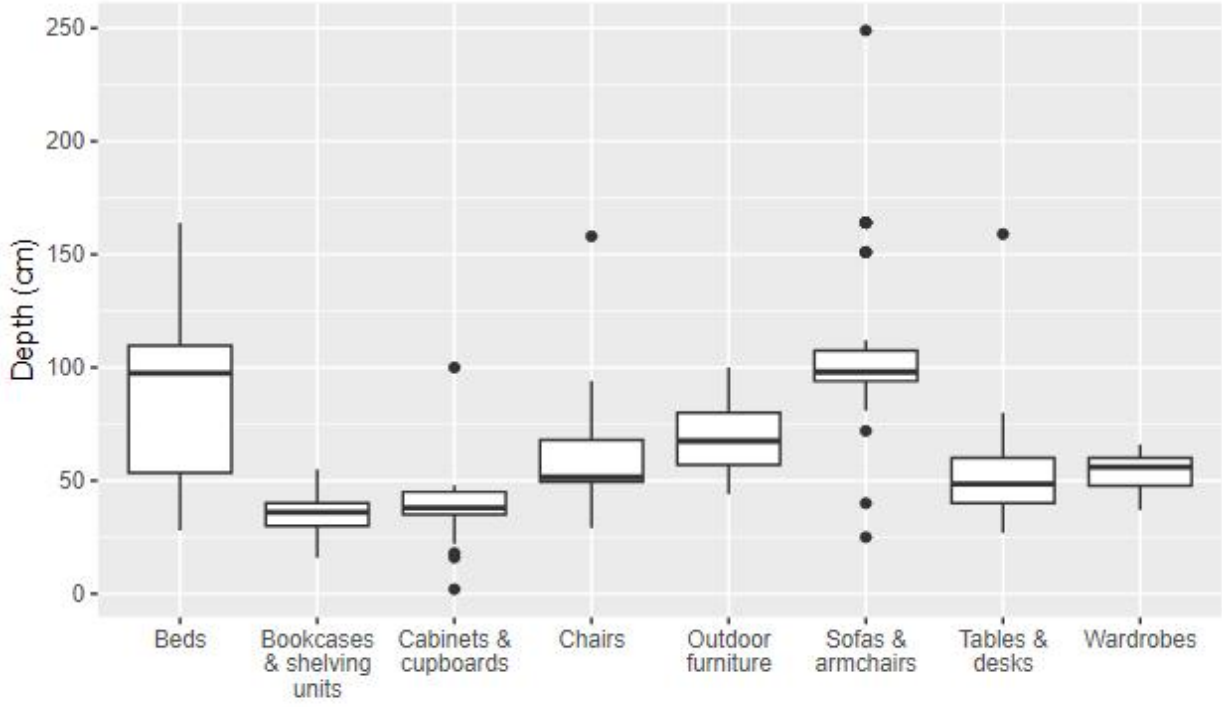
# MISSING DATA

- The missing values only exist in the height, depth and width variables.
- Dropping all rows with a missing value would lose too much data.
- Find a proper way to deal with the missing values.
- Consider replacing missing values with the mean or the median.

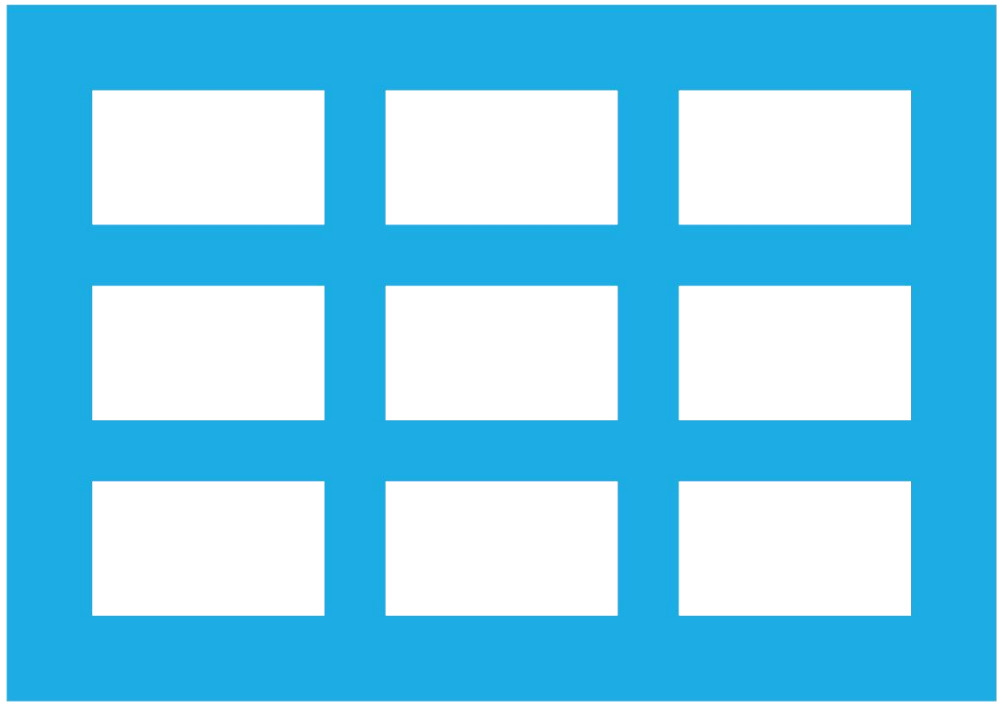
| Skim variable | Number of Missing value | Complete rate |
|---|---|---|
| Depth | 199 | 0.602 |
| Height | 131 | 0.738 |
| Width | 64 | 0.872 |

# MISSING DATA

- The type of items should affect its size, check the distributions for each category.
- Distributions vary a lot between categories, and often have outliers.
- Use the median instead of the mean to replace the missing value.
- Remove data that are missing in all three dimensions.



# MODEL ANALYSIS



# FIRST MODEL

- Two of sixteen categories are showed here, and none of them are significant.
- Sellable online, other colors and depth are also insignificant.
- Build another model without categories to examine whether they are significant.
- The AIC of this regression model is 359.3.

| | Parameter Estimate | 95% CI Lower Bound | 95% CI Upper Bound |
|---|---|---|---|
| Category Beds | 15.072 | -2630.028 | 2660.171 |
| Category Chairs | 16.171 | -2628.928 | 2661.270 |
| Sellable Online | 14.739 | -2775.501 | 2804.979 |
| Other Colors | 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 |

# SECOND MODEL

- Sellable online and other colors are still not significant, however depth is significant.
- The significant parameters here are depth, height, width.
- Build the third model with three significant factors.
- The AIC of this regression model is 398.8.

| | Parameter Estimate | 95% CI Lower Bound | 95% CI Upper Bound |
|---|---|---|---|
| Sellable Online | 13.056 | -1213.922 | 1240.034 |
| Other Colors | 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 |

# THIRD MODEL

- Build the third multiple logistic regression model of depth, height and width.
- All parameters are significant now.
- The AIC of the third model is 395.6 which is less than the former one(398.8).

| | 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 |

# THIRD MODEL

- For every extra centimetre in depth, we expect the log-odds to increase by 0.022
- Equivalent to multiplying the odds by 1.023.
- For width, log-odds increases per centimetre by 0.019 (odds multiplier of 1.019).
- For height, log-odds increases per centimetre by 0.007 (odds multiplier of 1.007).

| | 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 |

# CONCLUSION

- Bigger furniture (high width, depth and height size) seem to be more likely to have a selling price over 1,000 Saudi Riyals.
- The depth has the biggest impact, the width has the next biggest, followed by height.
- For further work, investigate the items' volume as a variable.
- We could also look further into the categories.

# THE END

Thank you!

Any questions?