Impact of furniture properties on whether people want to spend more than 1,000 Saudi Riyals

Group 13

Jingxuan Peng Ningqi Zhang Hancheng Lu

Sian Turner Ajay Sreekumar Kuzhuvelil

Introduction

- This dataset contains some information on furniture available for purchase from IKEA in Saudi Arabia. Due to the specification, we chose the items whose price is higher than 1000 Saudi Riyals for our study.

- The aim of our analysis is to ***evaluate the influence of furniture properties on whether they cost more than 1000 Saudi Riyals by using Generalized Linear Model (GLM).***

Exploratory data analysis

Data summarization

```
       X              item_id          category            price       
 Min.   :  1.0   Min.   :   91415   Length:500         Min.   :   3.0  
 1st Qu.:125.8   1st Qu.:20344743   Class :character   1st Qu.: 168.8  
 Median :250.5   Median :49284139   Mode  :character   Median : 457.0  
 Mean   :250.5   Mean   :48073026                      Mean   : 991.1  
 3rd Qu.:375.2   3rd Qu.:70414743                      3rd Qu.:1245.0  
 Max.   :500.0   Max.   :99305158                      Max.   :8551.0  

 sellable_online other_colors          depth            height      
 Mode :logical   Length:500         Min.   :  1.00   Min.   :  3.0  
 FALSE:1         Class :character   1st Qu.: 37.00   1st Qu.: 68.0  
 TRUE :499       Mode  :character   Median : 46.00   Median : 83.0  
                                    Mean   : 53.34   Mean   :102.3  
                                    3rd Qu.: 60.00   3rd Qu.:123.8  
                                    Max.   :252.00   Max.   :251.0  
                                    NA's   :191      NA's   :146    

     width        thousandRiyals    
 Min.   :  2.0   Length:500         
 1st Qu.: 56.0   Class :character   
 Median : 80.0   Mode  :character   
 Mean   :101.1                      
 3rd Qu.:134.2                      
 Max.   :367.0                      
 NA's   :80                         
```

standard five-number summary of variables (lots of NA)

Exploratory data analysis

Data cleaning

A tibble: 500 x 10 Groups: category [17]

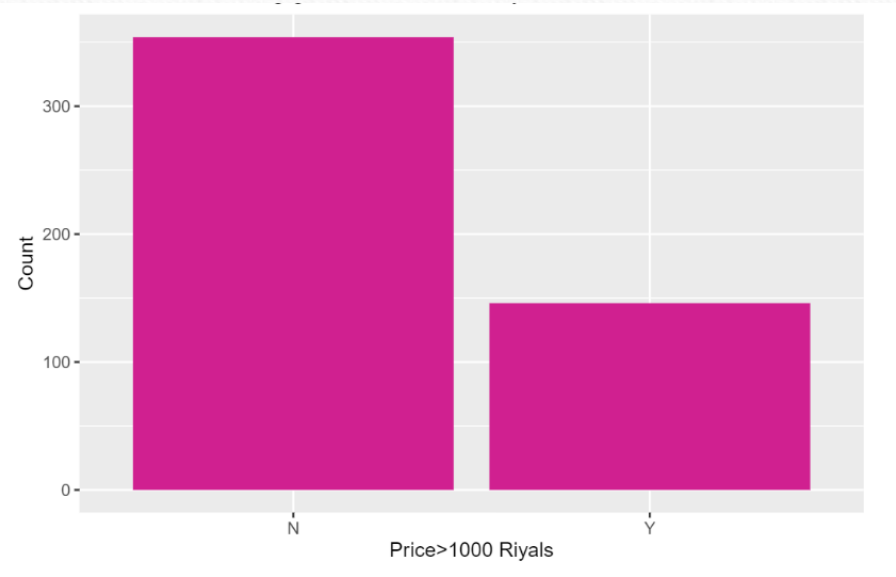
X <int>	item_id <int>	category <chr>	price <dbl>	sellable_online <lgl>	other_colors <chr>	depth <dbl>	height <dbl>	width <dbl>	thousandRiyals <chr>
1	60406785	Chairs	149.0	TRUE	No	44.00000	103.00000	52.00000	N
2	99221446	Tables & desks	5465.0	TRUE	No	50.45455	65.65217	74.10909	Y
3	29221765	Bookcases & shelving units	660.0	TRUE	No	35.00000	176.00000	86.00000	N
4	20403374	Sofas & armchairs	395.0	TRUE	Yes	106.92593	76.59375	161.63636	N
5	89256262	Chairs	1895.0	TRUE	Yes	164.00000	104.00000	117.00000	Y
6	50165079	Nursery furniture	375.0	TRUE	No	53.00000	93.00000	52.00000	N
7	40284191	Sofas & armchairs	150.0	TRUE	No	106.92593	15.00000	161.63636	N
8	19291315	Cabinets & cupboards	375.0	TRUE	No	47.00000	92.00000	45.00000	N
9	439468	Wardrobes	995.0	TRUE	No	47.00000	186.00000	120.00000	N
10	49252553	Chairs	400.0	TRUE	No	61.00000	87.00000	56.00000	N

1-10 of 500 rows Previous 1 2 3 4 5 6 ... 50 Next

final summary (thousandRiyals & mean replacement)

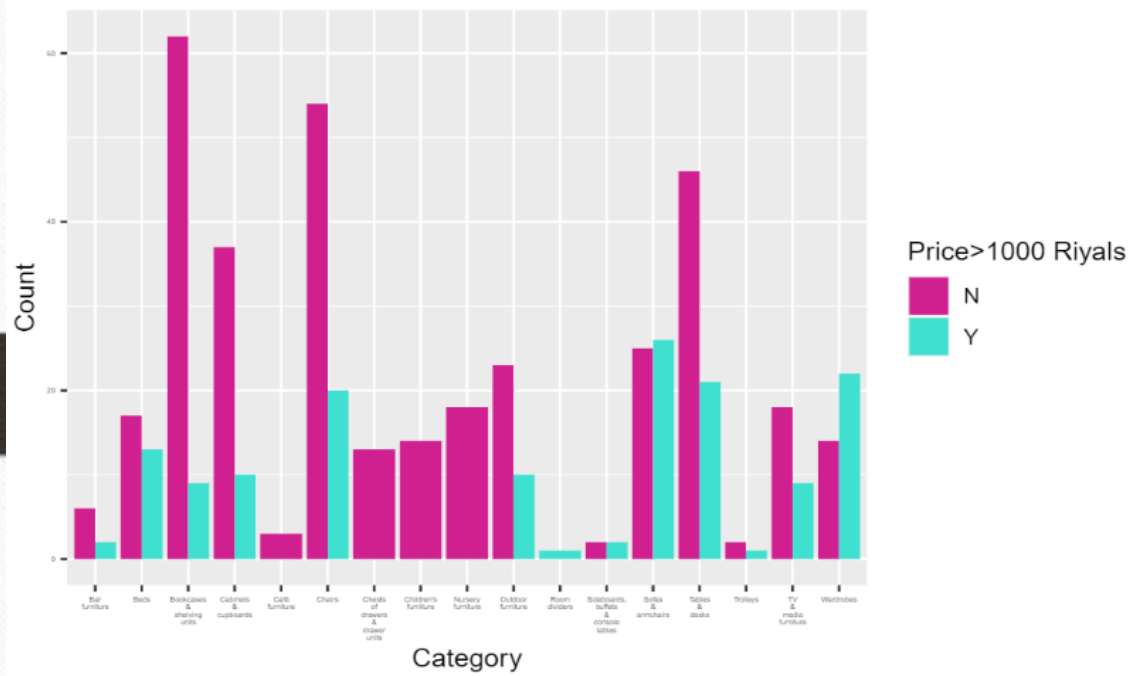
Exploratory data analysis

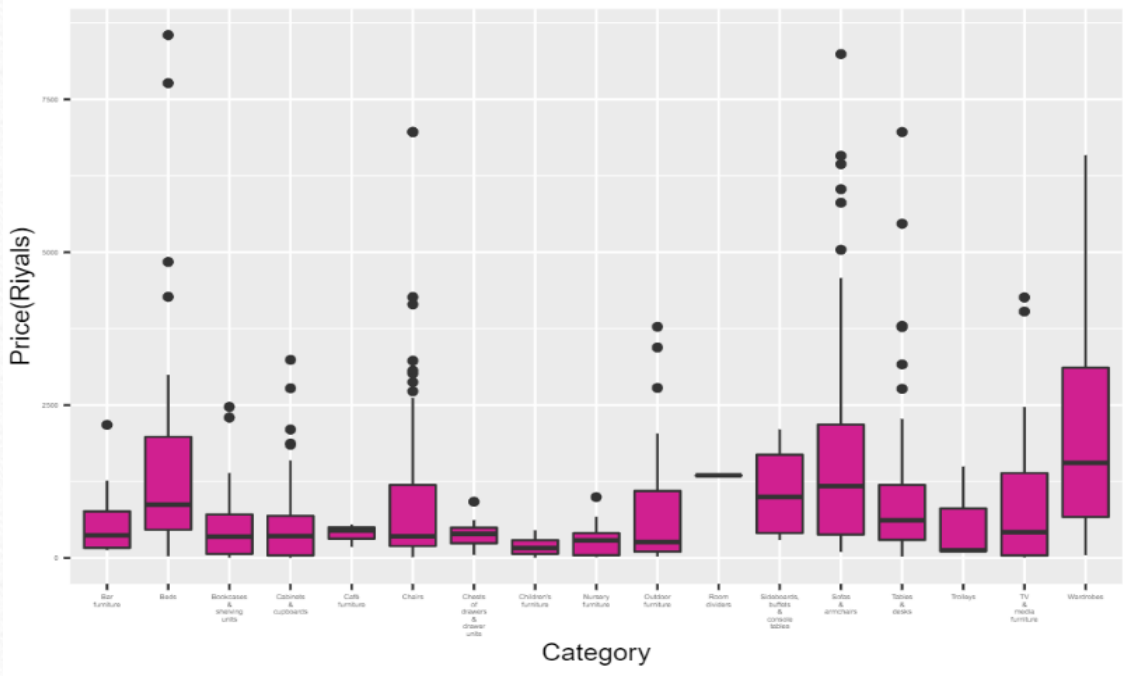
Visualization



Number of furniture with price greater than 1000 SR

The relationship between the category and the price of item



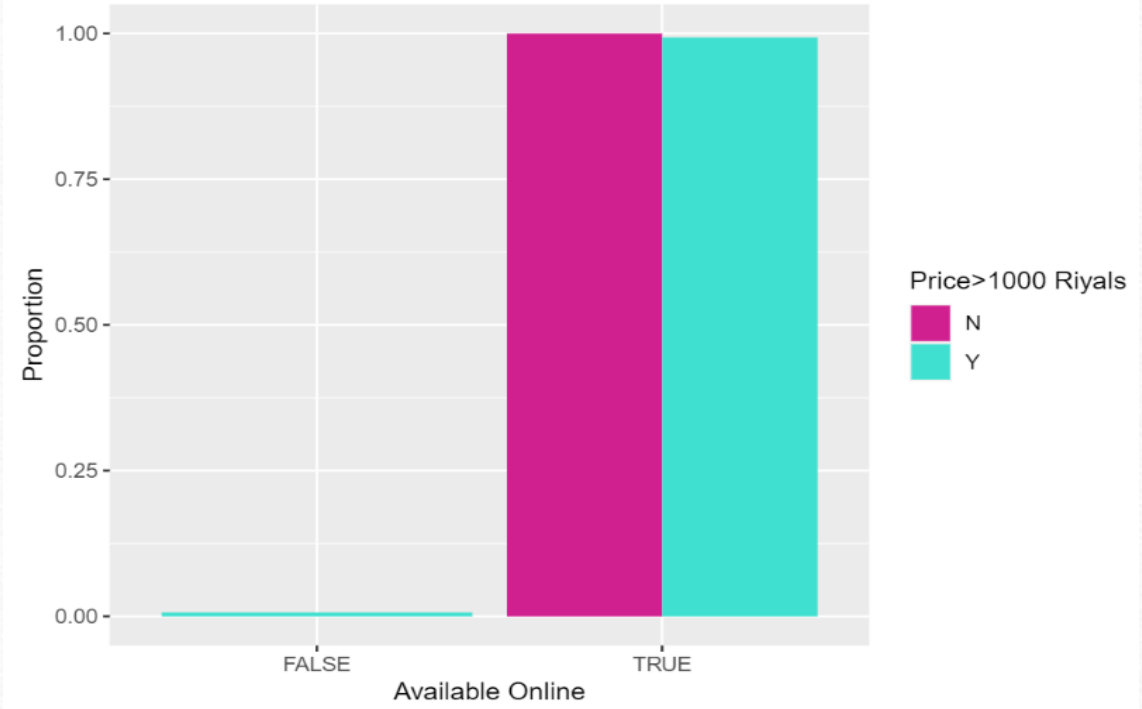


Total number in each category by price (sofas & armchairs)

Potential outlier in some categories (wardrobes)

The relationship between the sellable online and the price of item

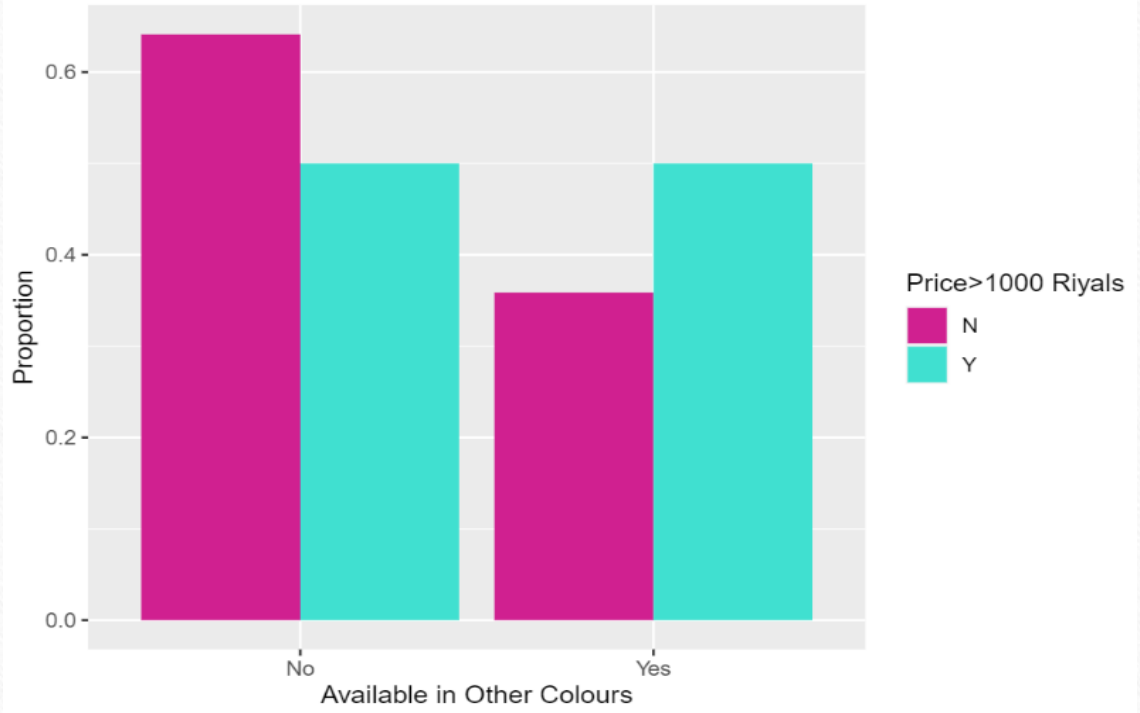
thousandRiyals	FALSE	TRUE
N	0.0% (0)	100.0% (354)
Y	0.7% (1)	99.3% (145)



The proportion of items available online and price greater than 1000 SR (145)

The relationship between other colors and the price of item

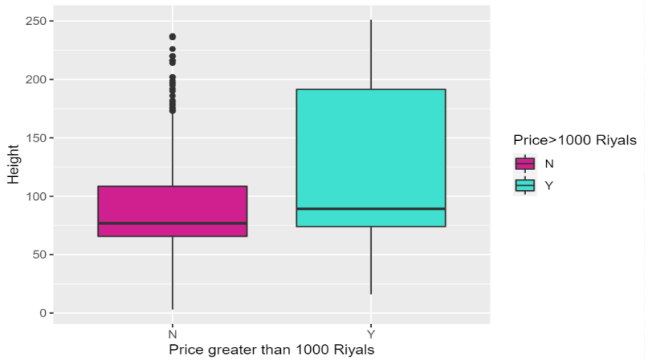
thousandRiyals	No	Yes
N	64.1% (227)	35.9% (127)
Y	50.0% (73)	50.0% (73)

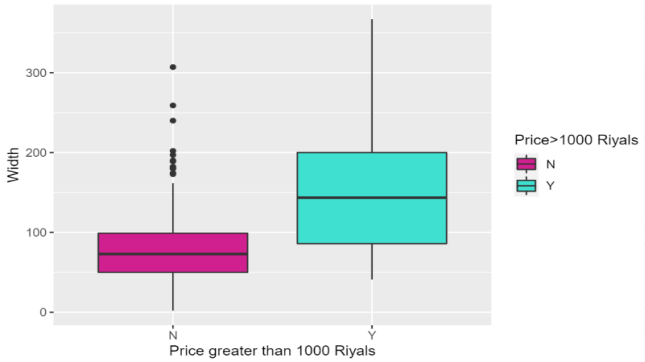


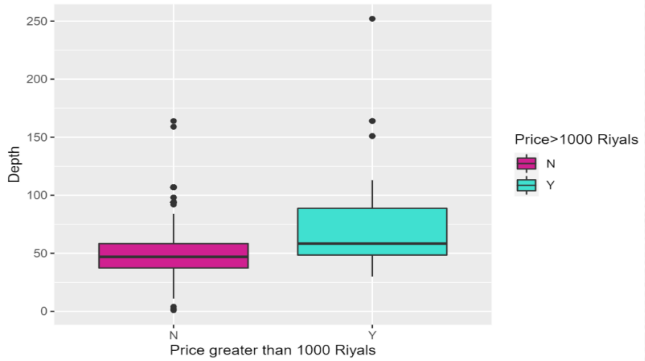
The proportion of items in different colors and price greater than 1000 SR (73)

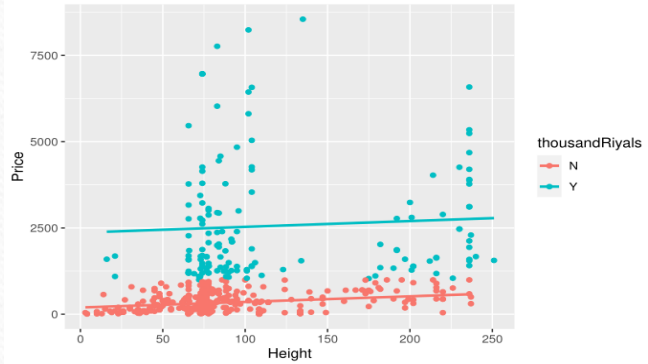
The relationship between size and the price of item

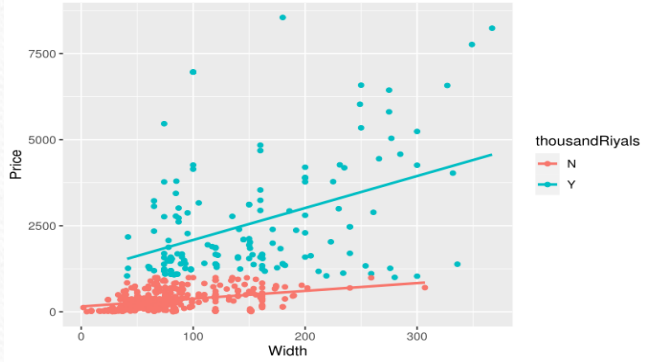
height, width and depth of item by the price

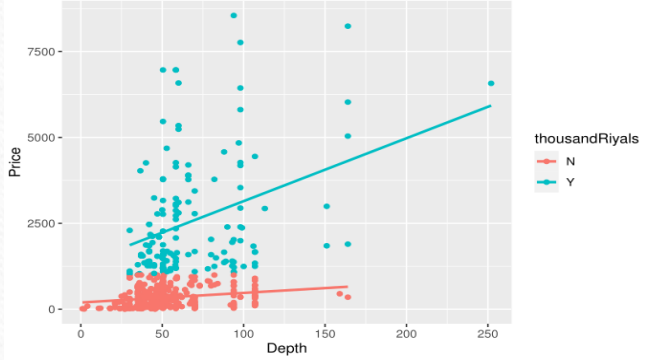












size against price (positive)

Statistical modelling & results

Methods

- Generalised Linear Models
- Multiple variables – step function (AIC values)

Steps

- Selection of input variables (Omit)
- Logit link function / Probit link function / Log-log link function

Selection of variables

Analysis of sellable_online and item_id variables

- sellable_online

sellable_online	N	Y
FALSE	0.0% (0)	100.0% (1)
TRUE	70.9% (354)	29.1% (145)

```
##
## Call:
## glm(formula = thousandRiyals ~ sellable_online, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8286  -0.8286  -0.8286   1.5722   1.5722
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)            13.57     535.41   0.025    0.980
## sellable_onlineTRUE   -14.46     535.41  -0.027    0.978
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 603.93  on 499  degrees of freedom
## Residual deviance: 601.47  on 498  degrees of freedom
## AIC: 605.47
##
## Number of Fisher Scoring iterations: 12
```

- item_id

only gives the unique identity of the item and does not have statistical significance.

- different combinations of the remaining variables

Fit GLM by using logit link function

- Fitted model

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{category} + \gamma \cdot \text{height} + \delta \cdot \text{width}$$

$$= -4.878 - 2.94 \cdot \mathbb{I}_{\text{category}}\ \text{(Beds)} - 7.349 \cdot \mathbb{I}_{\text{category}}\ \text{(Bookcases and shelving units)} - 2.894 \cdot \mathbb{I}_{\text{category}}\ \text{(Cabinets and cupboards)}$$
$$- 17.61 \cdot \mathbb{I}_{\text{category}}\ \text{(Cafe furniture)} - 0.4754 \cdot \mathbb{I}_{\text{category}}\ \text{(Chairs)} - 17.95 \cdot \mathbb{I}_{\text{category}}\ \text{(Chests of drawers and rawer units)}$$
$$- 18.48 \cdot \mathbb{I}_{\text{category}}\ \text{(Children's furniture)} - 18.41 \cdot \mathbb{I}_{\text{category}}\ \text{(Nursery furniture)} - 0.8598 \cdot \mathbb{I}_{\text{category}}\ \text{(Outdoor furniture)}$$
$$+ 12.52 \cdot \mathbb{I}_{\text{category}}\ \text{(Room dividers)} - 1.86 \cdot \mathbb{I}_{\text{category}}\ \text{(RSideboards, buffets and console tables)} - 2.525 \cdot \mathbb{I}_{\text{category}}\ \text{(Sofas and armchairs)}$$
$$- 0.2582 \cdot \mathbb{I}_{\text{category}}\ \text{(Tables and desks)} + 0.4862 \cdot \mathbb{I}_{\text{category}}\ \text{(Trolleys)} - 4.238 \cdot \mathbb{I}_{\text{category}}\ \text{(TV and media furniture)}$$
$$- 4.691 \cdot \mathbb{I}_{\text{category}}\ \text{(Wardrobes)} + 0.0248 \cdot \text{height} + 0.03524 \cdot \text{width}$$

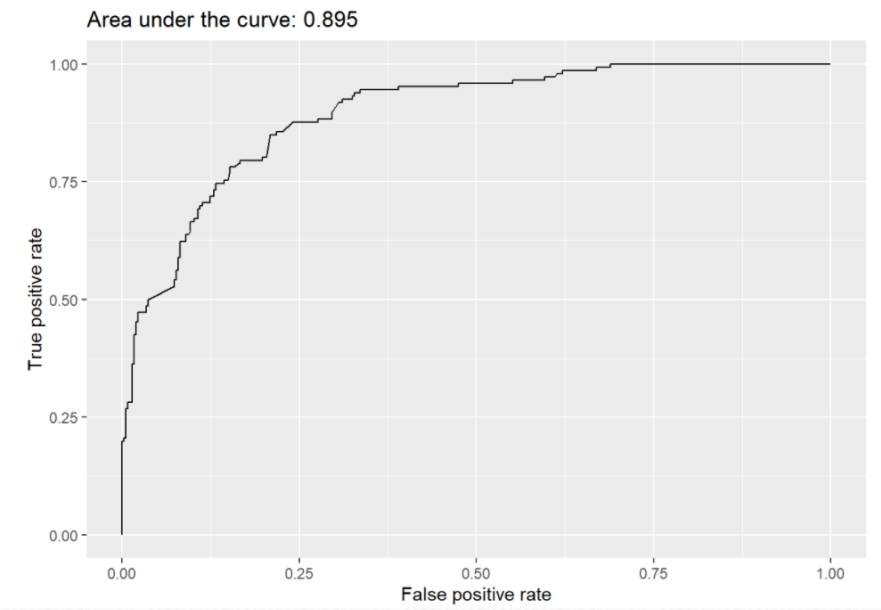
```
## Call:
## glm(formula = thousandRiyals ~ category + height + width, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4036  -0.6753  -0.1350   0.2557   2.9258
##
## Coefficients:
##                                              Estimate Std. Error z value  Pr(>|z|)
## (Intercept)                                -4.878e+00  9.979e-01  -4.888  1.02e-06 ***
## categoryBeds                               -2.940e+00  1.083e+00  -2.714  0.006648 **
## categoryBookcases & shelving units         -7.349e+00  1.358e+00  -5.414  6.18e-08 ***
## categoryCabinets & cupboards               -2.894e+00  1.026e+00  -2.821  0.004787 **
## categoryCaf?furniture                      -1.761e+01  3.736e+03  -0.005  0.996238
## categoryChairs                             -4.754e-01  8.825e-01  -0.539  0.590098
## categoryChests of drawers & drawer units   -1.795e+01  1.663e+03  -0.011  0.991389
## categoryChildren's furniture               -1.848e+01  1.504e+03  -0.012  0.990196
## categoryNursery furniture                  -1.841e+01  1.393e+03  -0.013  0.989458
## categoryOutdoor furniture                  -8.598e-01  9.637e-01  -0.892  0.372339
## categoryRoom dividers                       1.252e+01  6.523e+03   0.002  0.998469
## categorySideboards, buffets & console tables -1.860e+00  1.602e+00  -1.161  0.245717
## categorySofas & armchairs                  -2.525e+00  1.043e+00  -2.422  0.015434 *
## categoryTables & desks                     -2.582e-01  8.931e-01  -0.289  0.772532
## categoryTrolleys                            4.862e-01  1.609e+00   0.302  0.762468
## categoryTV & media furniture               -4.238e+00  1.196e+00  -3.542  0.000397 ***
## categoryWardrobes                          -4.691e+00  1.269e+00  -3.696  0.000219 ***
## height                                      2.480e-02  5.224e-03   4.747  2.07e-06 ***
## width                                       3.524e-02  4.659e-03   7.563  3.94e-14 ***
## ---
## 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: 361.11  on 481  degrees of freedom
## AIC: 399.11
##
## Number of Fisher Scoring iterations: 17
```

Performance of model

- Hosmer and Lemeshow Test

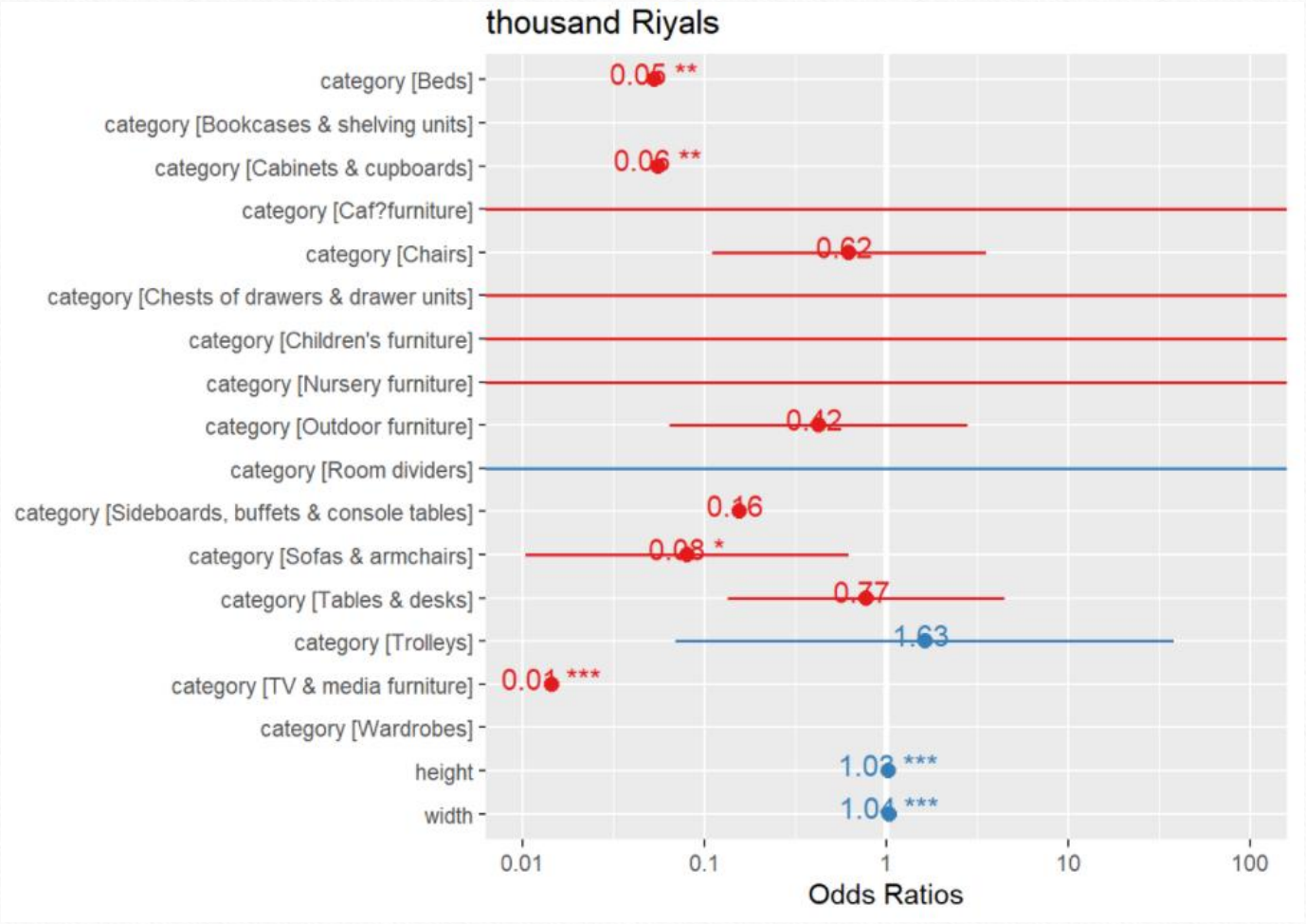
```
## Hosmer and Lemeshow Goodness-of-Fit Test
##
## Call:
## glm(formula = thousandRiyals ~ category + height + width, family = binomial(link = "logit"),
##     data = data)
##  ChiSquare df   P_value
##   7.129942  6 0.3089973
```

- ROC curve and AUC for the model fitted

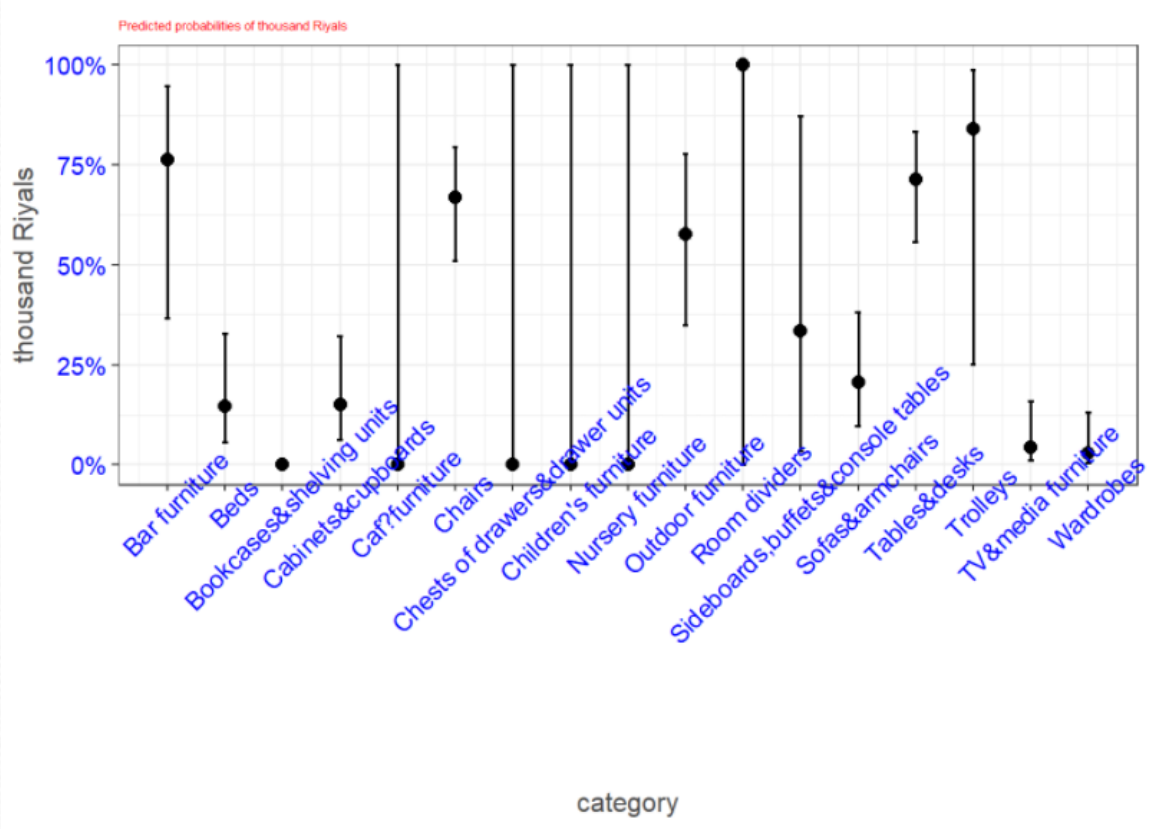


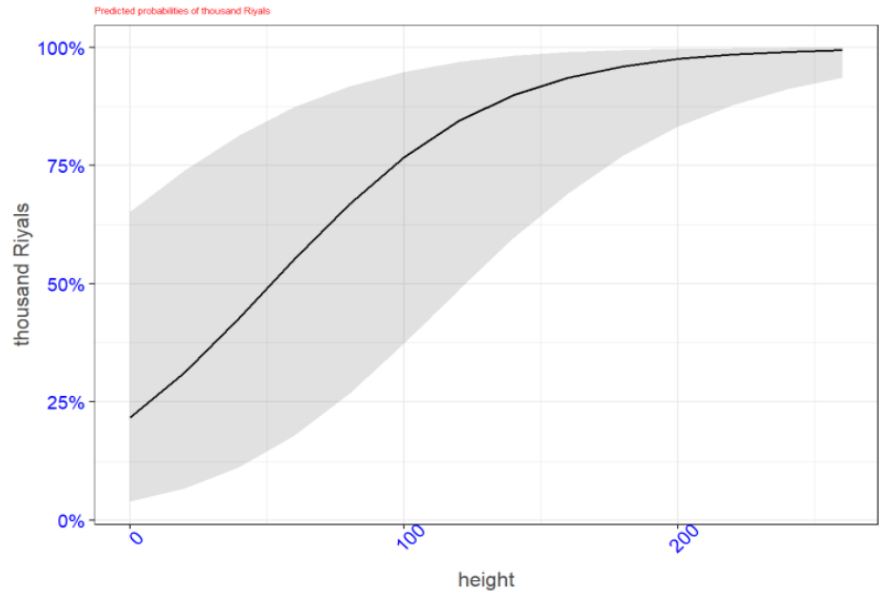
quantify the effect of each of the predictors

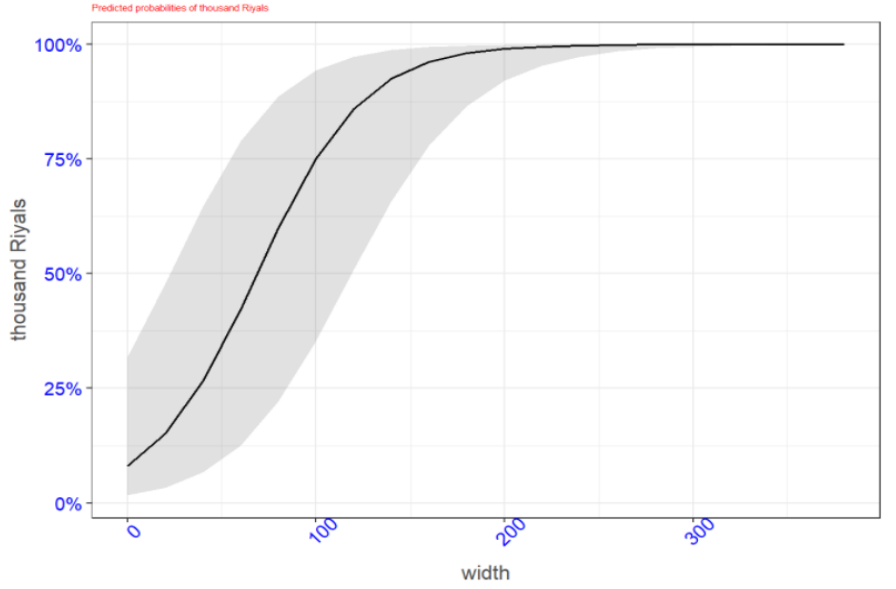
Odds ratios



Odds ratios of each of the predictors.







Fit GLM by using probit link function

```
##
## Call:
## glm(formula = thousandRiyals ~ category + height + width, family = binomial(link = "probit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3583  -0.6917  -0.1052   0.2445   2.9527
##
## Coefficients:
##                                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                -2.706e+00  5.657e-01  -4.784 1.72e-06 ***
## categoryBeds                               -1.442e+00  6.112e-01  -2.359 0.018302 *
## categoryBookcases & shelving units         -3.913e+00  7.291e-01  -5.367 8.02e-08 ***
## categoryCabinets & cupboards               -1.523e+00  5.778e-01  -2.636 0.008397 **
## categoryCaf?furniture                      -5.229e+00  8.895e+02  -0.006 0.995309
## categoryChairs                             -2.690e-01  5.169e-01  -0.520 0.602801
## categoryChests of drawers & drawer units   -5.821e+00  3.749e+02  -0.016 0.987613
## categoryChildren's furniture               -6.330e+00  3.322e+02  -0.019 0.984795
## categoryNursery furniture                  -6.171e+00  3.091e+02  -0.020 0.984073
## categoryOutdoor furniture                  -4.669e-01  5.613e-01  -0.832 0.405534
## categoryRoom dividers                       2.560e+00  1.569e+03   0.002 0.998698
## categorySideboards, buffets & console tables -8.915e-01  9.216e-01  -0.967 0.333346
## categorySofas & armchairs                  -1.225e+00  5.935e-01  -2.064 0.039033 *
## categoryTables & desks                     -1.288e-01  5.233e-01  -0.246 0.805532
## categoryTrolleys                            2.615e-01  9.360e-01   0.279 0.779935
## categoryTV & media furniture               -2.280e+00  6.714e-01  -3.395 0.000686 ***
## categoryWardrobes                          -2.436e+00  7.041e-01  -3.460 0.000540 ***
## height                                      1.339e-02  2.840e-03   4.715 2.42e-06 ***
## width                                       1.901e-02  2.366e-03   8.033 9.51e-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: 363.18  on 481  degrees of freedom
## AIC: 401.18
##
## Number of Fisher Scoring iterations: 17
```

- **Compare with logit link function**

  As the deviance value of this model is higher than the logit link function, the fit is better for the logit model.

Fit GLM by using log-log link function

```
##
## Call:
## glm(formula = thousandRiyals ~ category + height + width, family = binomial(link = "cloglog"),
##     data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.41988  -0.71977  -0.21201   0.04256   2.80570
##
## Coefficients:
##                                              Estimate Std. Error z value  Pr(>|z|)
## (Intercept)                                -4.088e+00  8.113e-01  -5.038  4.70e-07 ***
## categoryBeds                               -2.284e+00  8.859e-01  -2.578  0.009941 **
## categoryBookcases & shelving units         -5.462e+00  1.061e+00  -5.147  2.64e-07 ***
## categoryCabinets & cupboards               -2.060e+00  8.453e-01  -2.437  0.014809 *
## categoryCaf?furniture                      -1.728e+01  3.582e+03  -0.005  0.996151
## categoryChairs                             -1.675e-01  7.547e-01  -0.222  0.824351
## categoryChests of drawers & drawer units   -1.735e+01  1.652e+03  -0.011  0.991619
## categoryChildren's furniture               -1.766e+01  1.532e+03  -0.012  0.990803
## categoryNursery furniture                  -1.774e+01  1.386e+03  -0.013  0.989787
## categoryOutdoor furniture                  -4.724e-01  8.136e-01  -0.581  0.561484
## categoryRoom dividers                      -8.114e-01  6.591e+02  -0.001  0.999018
## categorySideboards, buffets & console tables -1.115e+00  1.127e+00  -0.989  0.322684
## categorySofas & armchairs                  -1.554e+00  8.351e-01  -1.860  0.062823 .
## categoryTables & desks                     -1.266e-01  7.636e-01  -0.166  0.868283
## categoryTrolleys                            6.177e-01  1.202e+00   0.514  0.607291
## categoryTV & media furniture               -2.806e+00  9.443e-01  -2.972  0.002963 **
## categoryWardrobes                          -3.345e+00  9.728e-01  -3.438  0.000585 ***
## height                                      1.973e-02  3.600e-03   5.481  4.23e-08 ***
## width                                       2.361e-02  3.034e-03   7.784  7.04e-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: 363.58  on 481  degrees of freedom
## AIC: 401.58
##
## Number of Fisher Scoring iterations: 17
```

- **Compare with previous link functions**

  As the deviance value of this model is higher than both logit and probit models, the loglog model fit is not better than any of the logit link function or the probit link function.

Conclusions

- Question: Which properties of furniture influence whether they cost more than 1000 Saudi Riyals?
- Response Variable: Price >1000, Yes; Price <1000, No.
- Explanatory Variable: sellable_online, category, other_colors, depth, height, width
- Method: Generalised Linear Model (GLM)
- Significant Variable: category, height, width
- Check the model: AUC=0.895 good
- Compare the other models: logit is better

Further work

- The variable “sellable_online” contains only one “FALSE” in the dataset, which is not representative. We could collect more data to identify this variable to assess if the online sales could affect the price.
- The variables “other_colors” and “depth” are not significant in this project, so we could collect more data to identify the influence of color and depth.
- More explanatory variables could be considered in the model.

Thank you!

2022.3.25