GLM

2023-03-18

Data Exploration

1

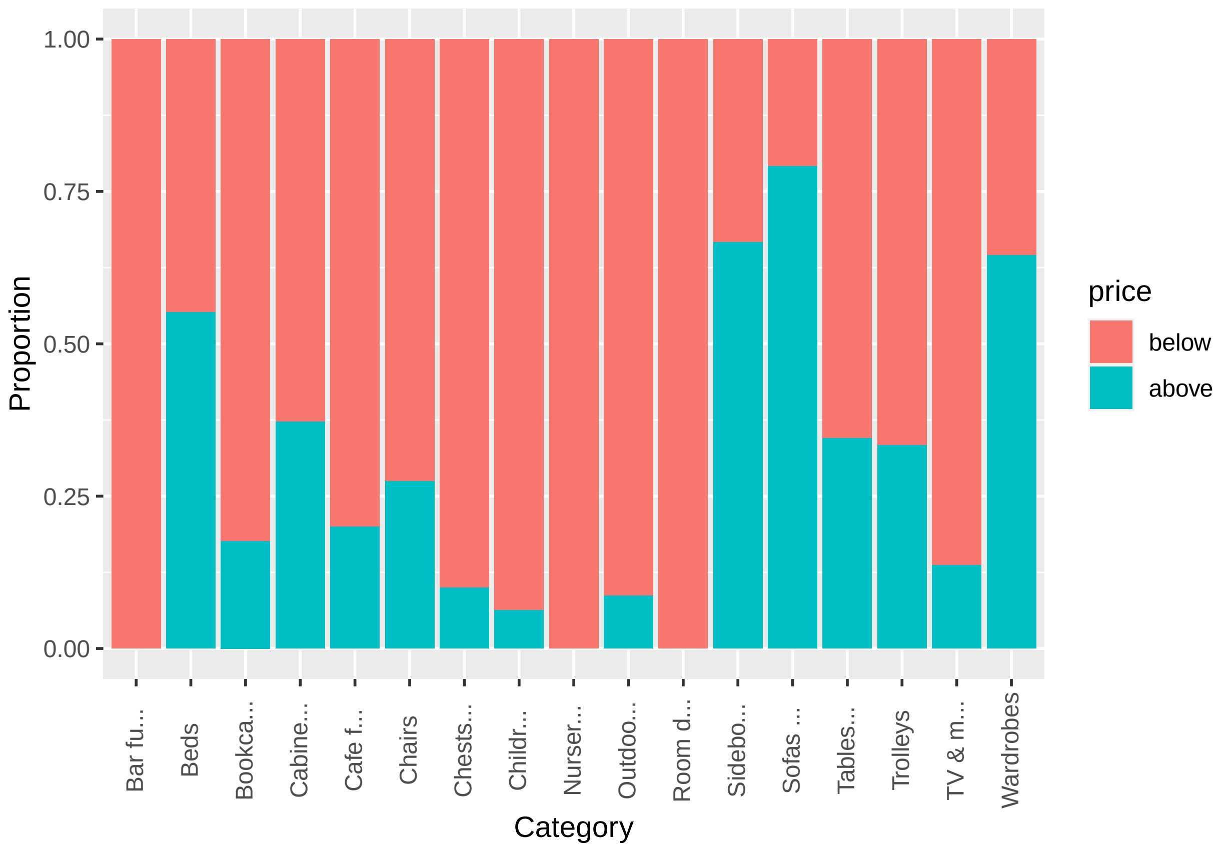
Create Response variable: Create a new variable indicating whether each item costs more than 1000 Saudi Riyals. Already done in the cleaning part

2

```r
furniture <- read.csv("cleaned_data.csv", stringsAsFactors = T)
furniture$price <- factor(furniture$price, levels = c(0, 1), labels = c("below", "above"))
```

The relationship category of price.

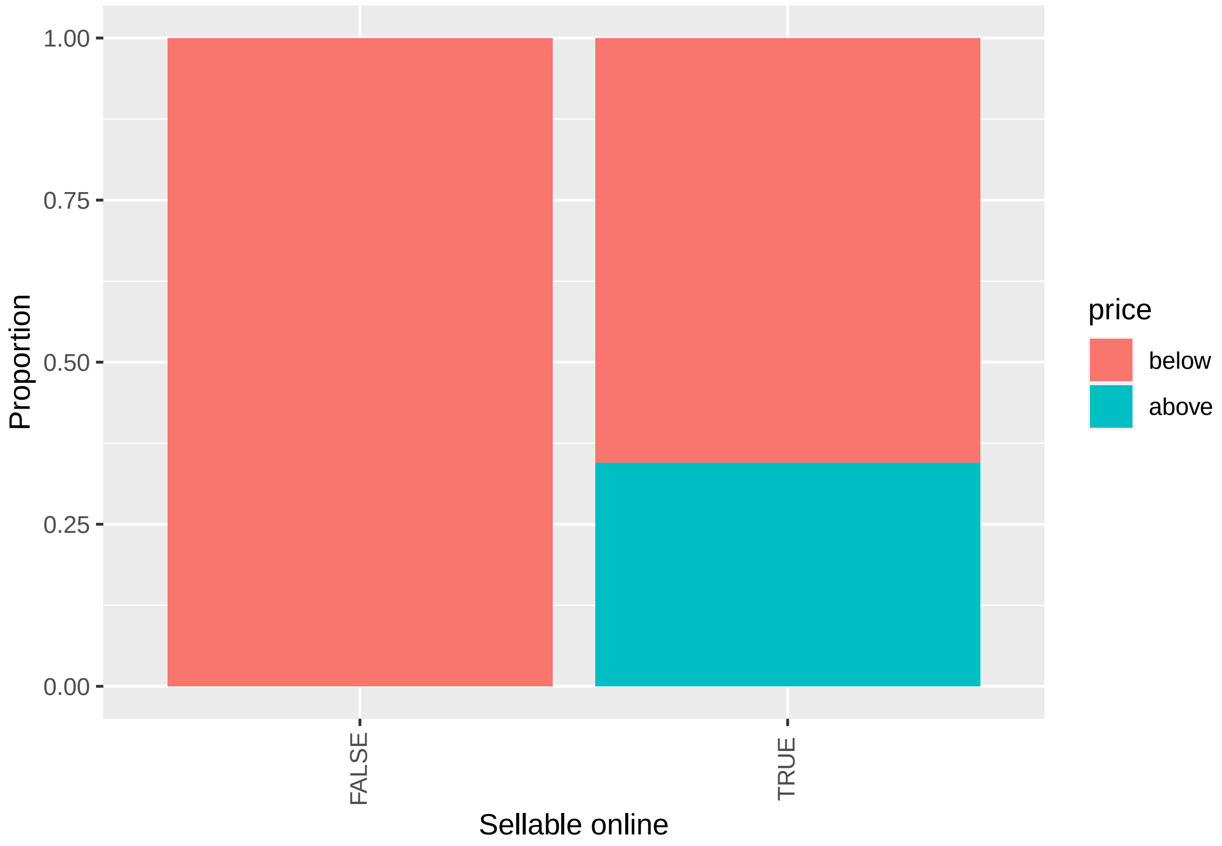
```r
library(ggplot2)
library(tidyverse)
furniture %>% ggplot(mapping=aes(x=str_trunc(as.character(category), 9, ell="..."), fill=price)) +
    geom_bar(position="fill") +
  theme(axis.text.x = element_text(angle = 90, vjust=0.4)) +
  xlab("Category") +
  ylab("Proportion")
```



Category of sofas and armchairs has the most proportion of the price above 1000.

Relationship between sellable_online and price.

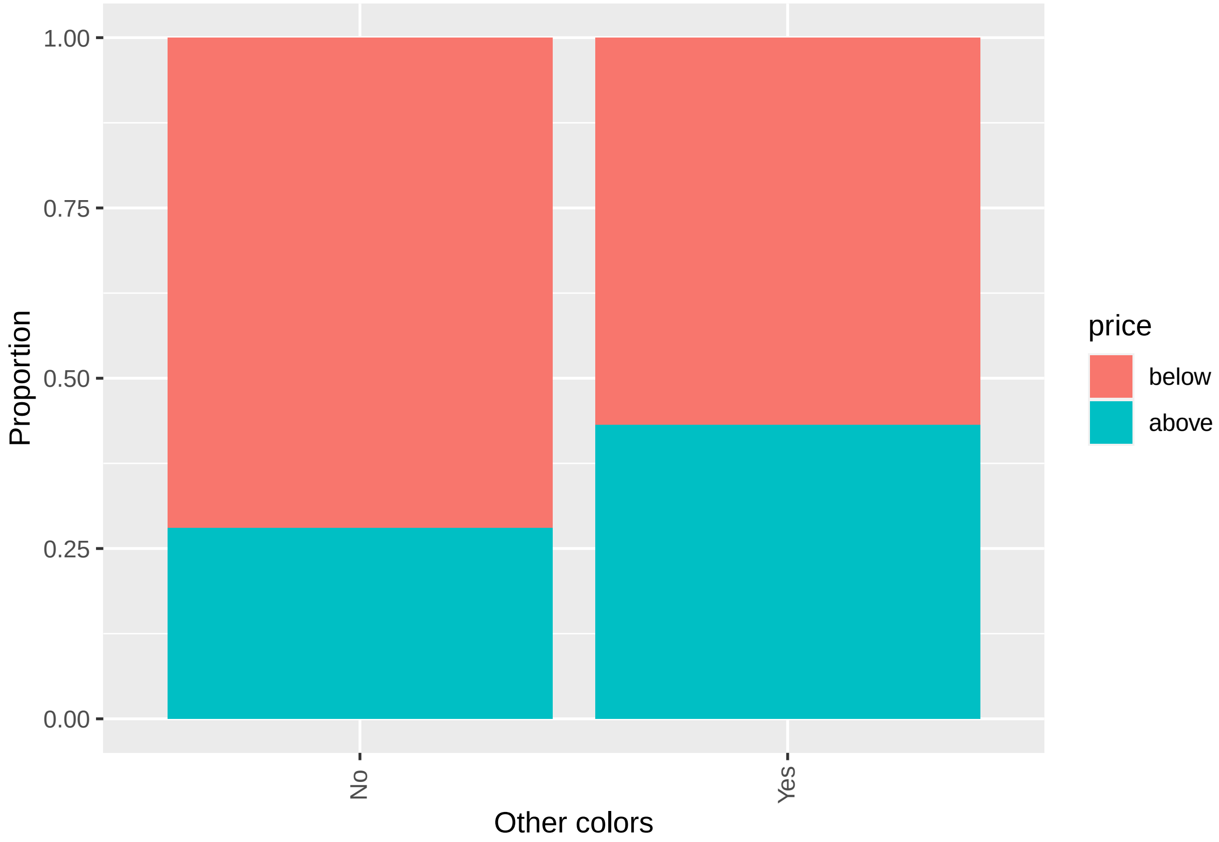
```
furniture %>% ggplot(mapping=aes(x=sellable_online, fill=price)) +
  geom_bar(position="fill") +
 theme(axis.text.x = element_text(angle = 90, vjust=0.4)) +
 xlab("Sellable online") +
 ylab("Proportion")
```



All unsellable online productions are under 1000

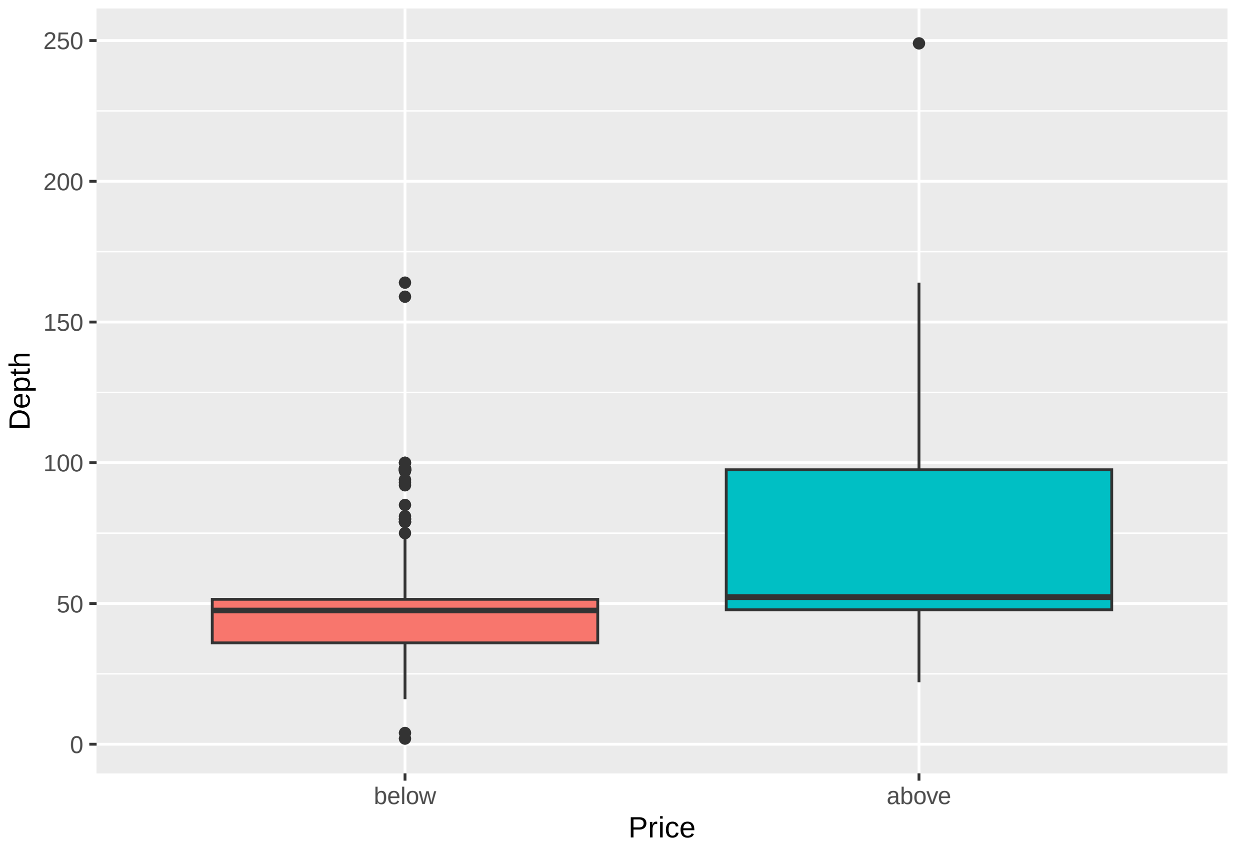
Relationship between other_colors and price.

```
furniture %>% ggplot(mapping=aes(x=other_colors, fill=price)) +
    geom_bar(position="fill") +
  theme(axis.text.x = element_text(angle = 90, vjust=0.4)) +
  xlab("Other colors") +
  ylab("Proportion")
```



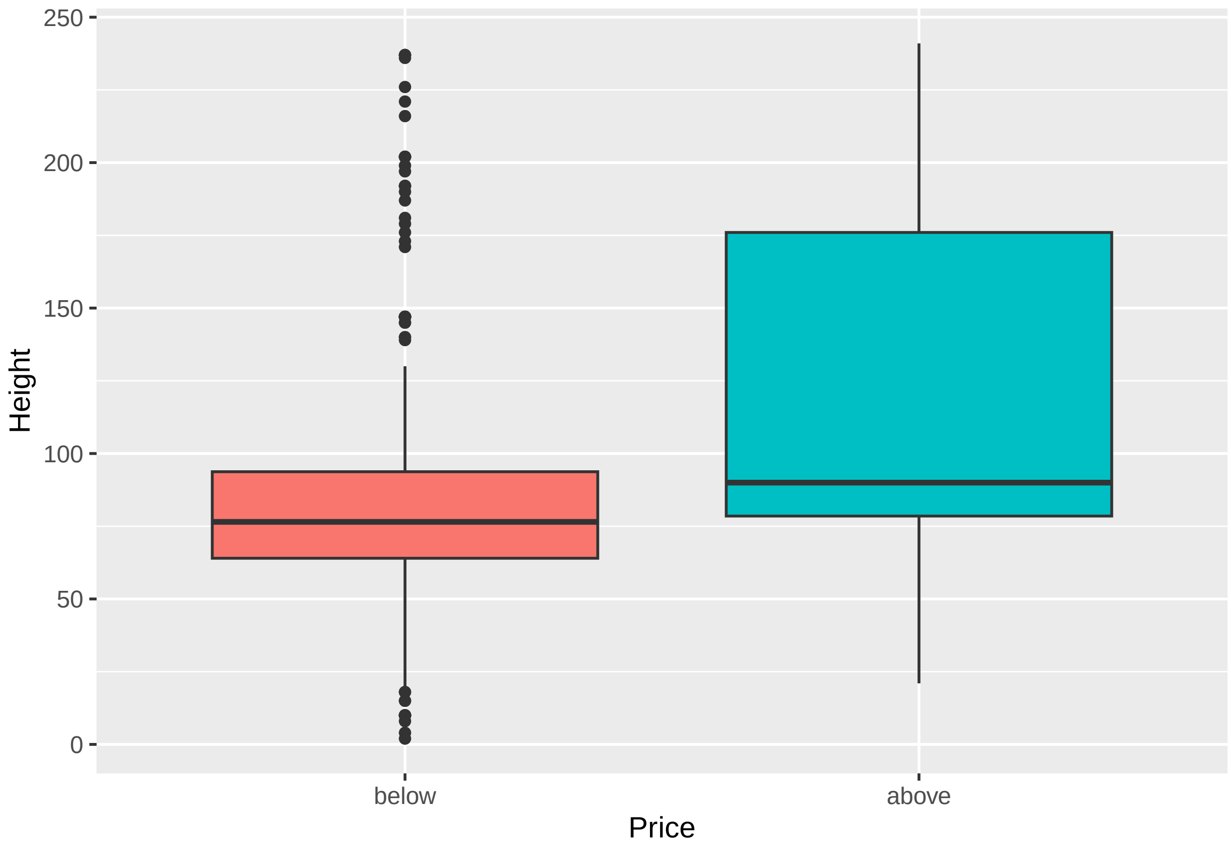
The proportion of above 1000 is higher for the other colors.

```
ggplot(furniture, aes(x = price, y = depth, fill = price)) +
  geom_boxplot() +
  labs(x = "Price", y = "Depth")+
  theme(legend.position = "none")
```

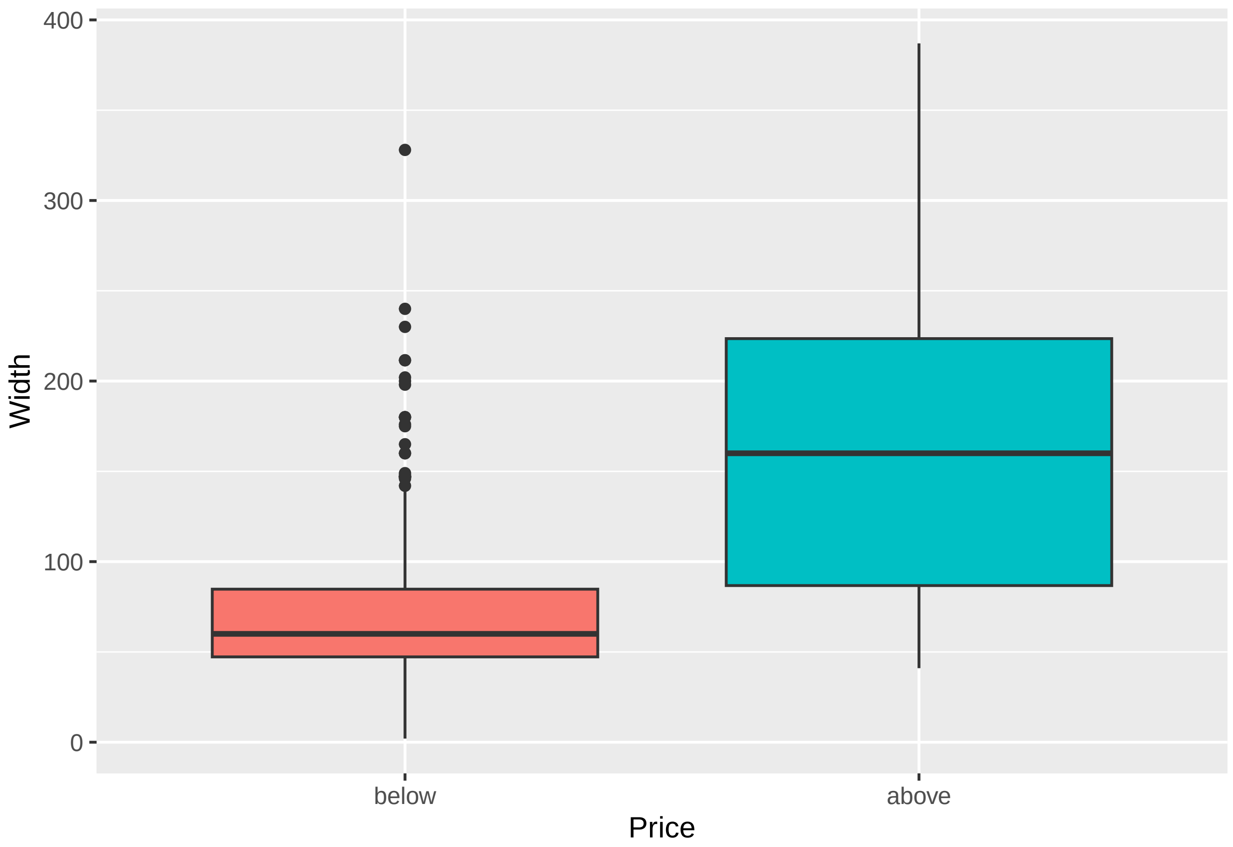


Depth between 50 and 100 seems like more possible to have price over1000

```
ggplot(furniture, aes(x = price, y = height, fill = price)) +
  geom_boxplot() +
  labs(x = "Price", y = "Height")+
  theme(legend.position = "none")
```



```
ggplot(furniture, aes(x = price, y = width, fill = price)) +
  geom_boxplot() +
  labs(x = "Price", y = "Width")+
  theme(legend.position = "none")
```



It Seems like the bigger the furniture is, the higher the price is

Modeling

Model 1

Build a multiple logistic regression model, use the sellable_one, other_colors, depth, height and width as the predictors, to predict the price.

```
# Fit a binary logistic regression model
model1 <- glm(price ~   category + sellable_online + other_colors + depth + height + width,
              data = furniture, family = binomial(link = "logit"))
coeffs <- summary(model1)$coefficients[, 1]
knitr::kable(cbind(coeffs, confint(model1)))
```

	coeffs	2.5 %	97.5 %
(Intercept)	-36.4522200	NA	114.4823701
categoryBeds	15.0716159	-101.8763844	NA
categoryBookcases & shelving units	10.8898688	-50.9498162	432.6639206
categoryCabinets & cupboards	14.6832516	-102.4487251	NA
categoryCafe furniture	16.2356891	-32.5365458	509.6395605
categoryChairs	16.1709871	-104.0320868	NA
categoryChests of drawers & drawer units	14.4495826	-31.5281949	523.1713267

	coeffs	2.5 %	97.5 %
categoryChildren's furniture	13.9816197	-29.0850783	538.6605844
categoryNursery furniture	-0.2738115	-20.2670075	18.9765743
categoryOutdoor furniture	14.0201206	-33.9292309	511.9345443
categoryRoom dividers	-3.2680703	-34.1252555	27.0214679
categorySideboards, buffets & console tables	16.1110762	-29.5469471	526.5858333
categorySofas & armchairs	15.5875315	-101.2211955	NA
categoryTables & desks	16.3748908	-104.2788392	NA
categoryTrolleys	15.7812653	-27.1676988	541.1057629
categoryTV & media furniture	14.2316896	-33.9174100	511.0512675
categoryWardrobes	12.3407649	-66.7926832	NA
sellable_onlineTRUE	14.7392460	-154.1074483	NA
other_colorsYes	0.0074583	-0.6108288	0.6120504
depth	0.0100914	-0.0075794	0.0283358
height	0.0284498	0.0179958	0.0404015
width	0.0265982	0.0196010	0.0344984

The sellable_onlineTRUE and other_colorsYes are not significant, because their p-values are larger than 0.05, while the depth, height and width are significant predictors here.

Model 2

Refit the model, using height, depth and width as the predictors:

```r
model2 <- glm(price~height+width+depth, furniture, family = "binomial")
coeffs <- summary(model2)$coefficients[, 1]
knitr::kable(cbind(coeffs, confint(model2)))
```

	coeffs	2.5 %	97.5 %
(Intercept)	-4.6997224	-5.6316281	-3.8414634
height	0.0070117	0.0020877	0.0120174
width	0.0189194	0.0141344	0.0241013
depth	0.0223073	0.0119229	0.0333362

The residual deviance of the second model is 387.63, which is only slightly higher than the residual deviance of the first model (386.80). However, the AIC value of the second model (395.63) is lower than the AIC value of the first model (398.8), indicating that the second model is a better fit for the data than the first model.

Overall, the second model suggests that the dimensions of the furniture (height, width, and depth) are the most important predictors of whether the price is more than 1000 Saudi Riyals, while the availability of online purchasing and other colors do not seem to have a significant impact on the furniture price. Therefore, the second model is a more parsimonious and interpretable model that could be used for predicting the price of furniture based on its dimensions.