DAS_Group23_GLM

2023-03-16

Data Exploration

1

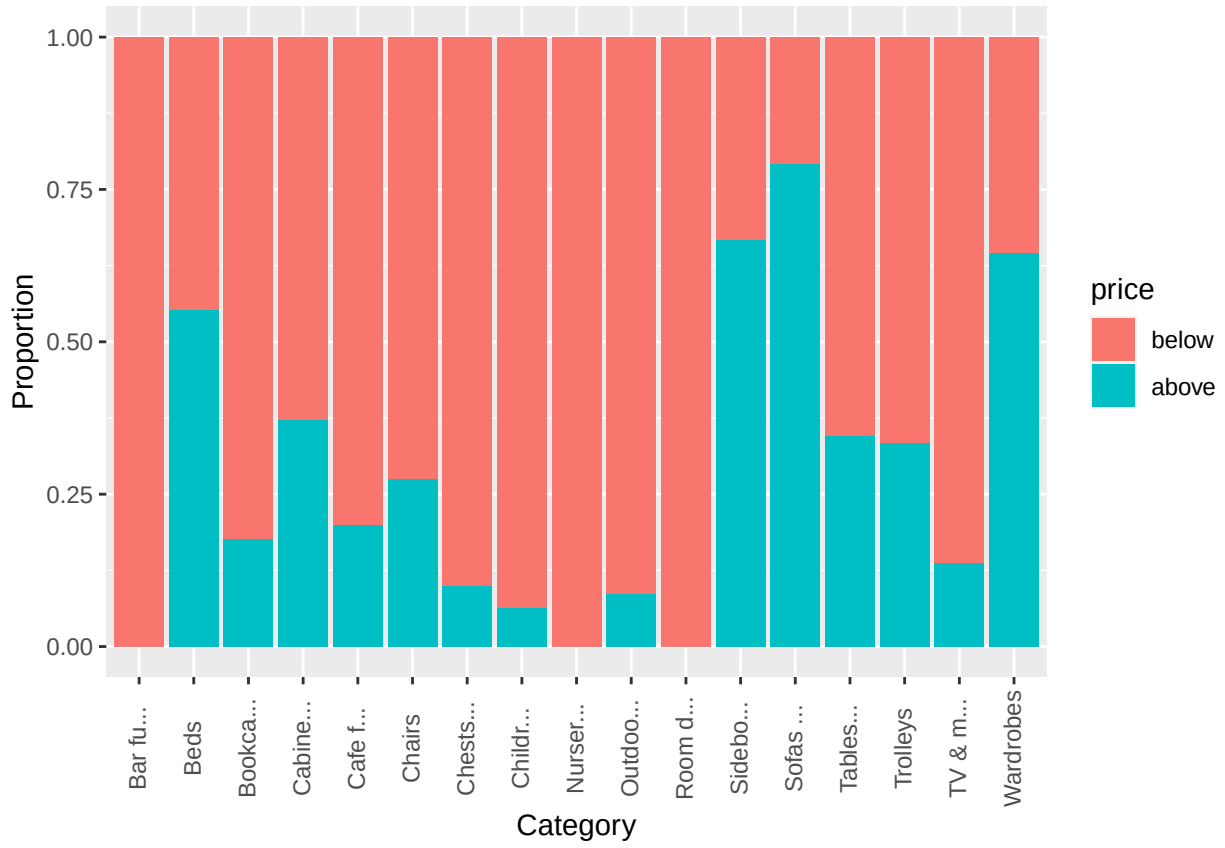
Create Response variable: Create a new variable indicating whether each item costs more than 1000 Saudi Riyals. Already done in the cleaning part

2

```
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.

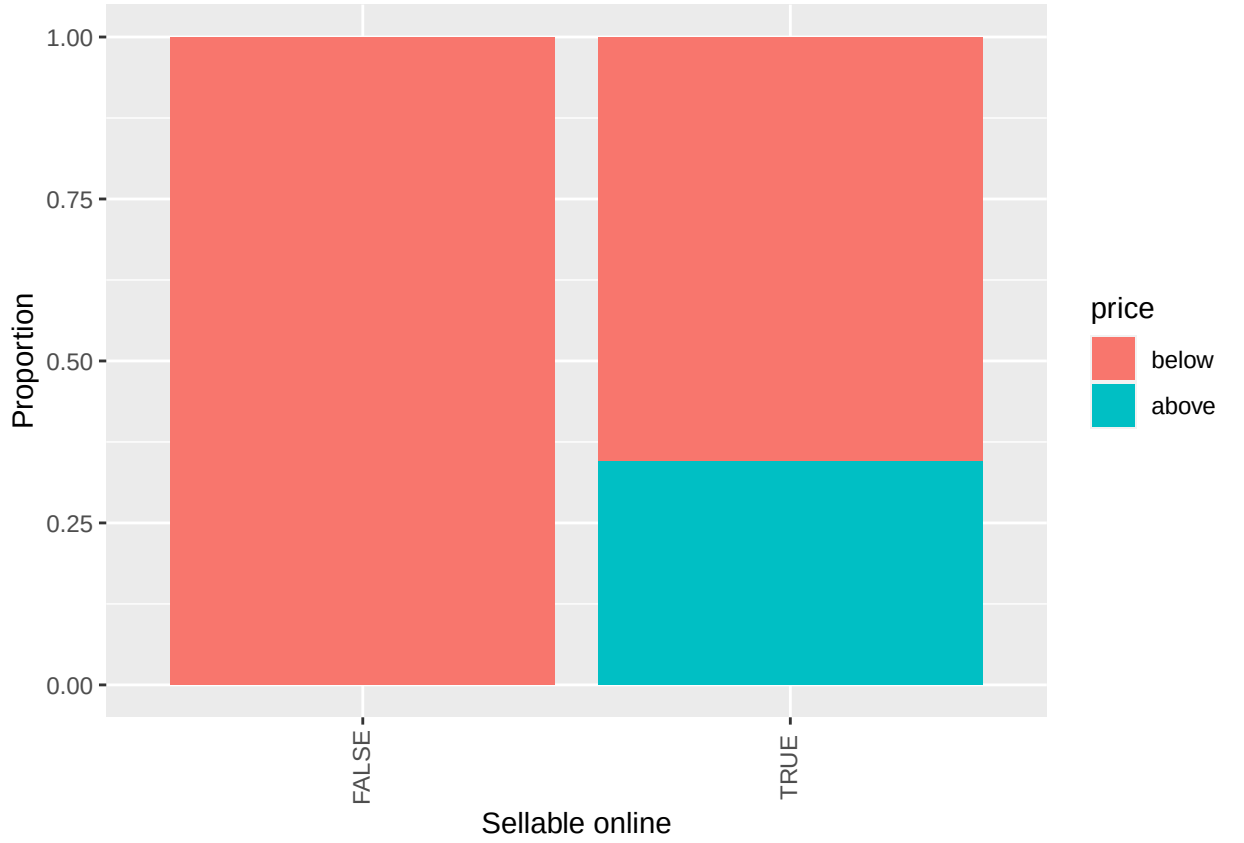
```
library(ggplot2)
library(tidyverse)
furniture %>% ggplot(mapping=aes(x=str_trunc(as.character(category), 9, ell="..."), fill=price)) +
    geom_bar(position="fill") +
  theme(axis.text.x = element_text(angle = 90, vjust=0.4)) +
  xlab("Category") +
  ylab("Proportion")
```



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

Relationship between sellable_online and price.

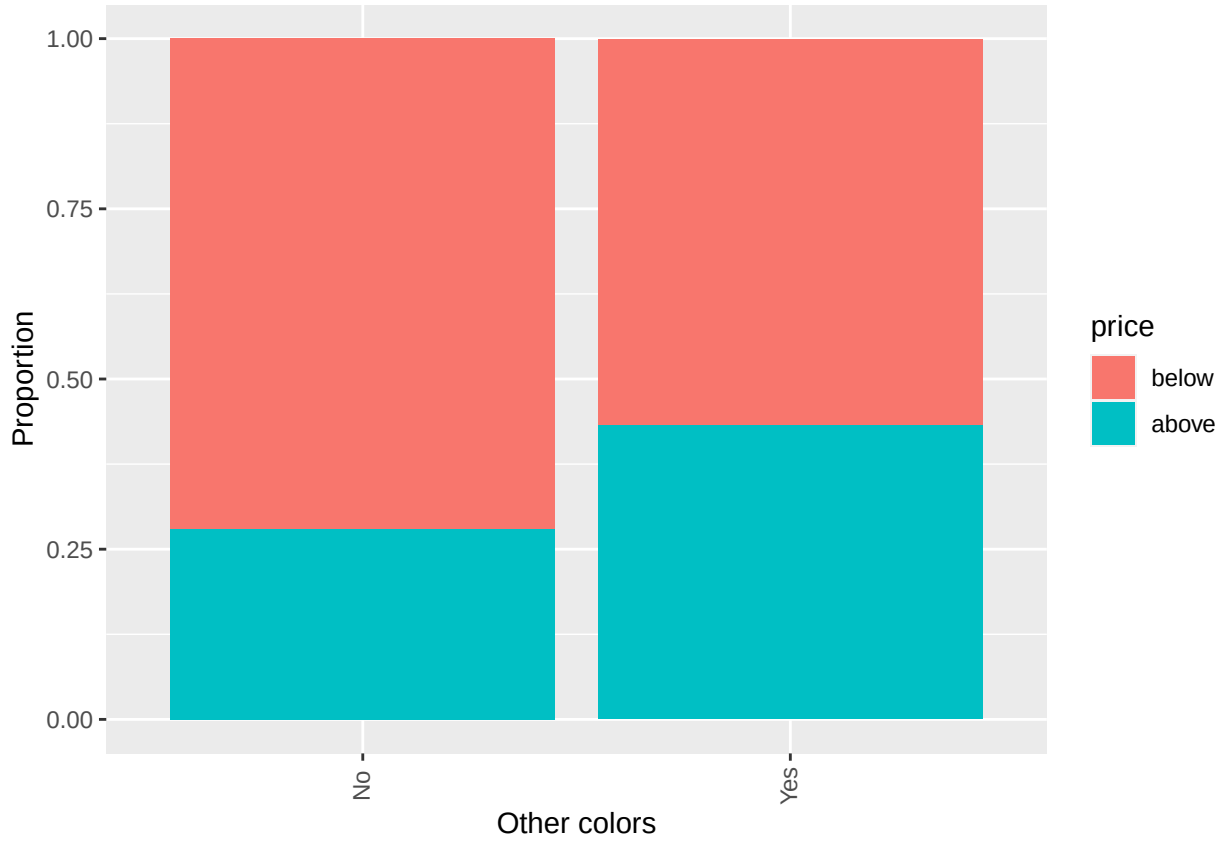
```
furniture %>% ggplot(mapping=aes(x=sellable_online, fill=price)) +
    geom_bar(position="fill") +
  theme(axis.text.x = element_text(angle = 90, vjust=0.4)) +
  xlab("Sellable online") +
  ylab("Proportion")
```



All unsellable online productions are under 1000

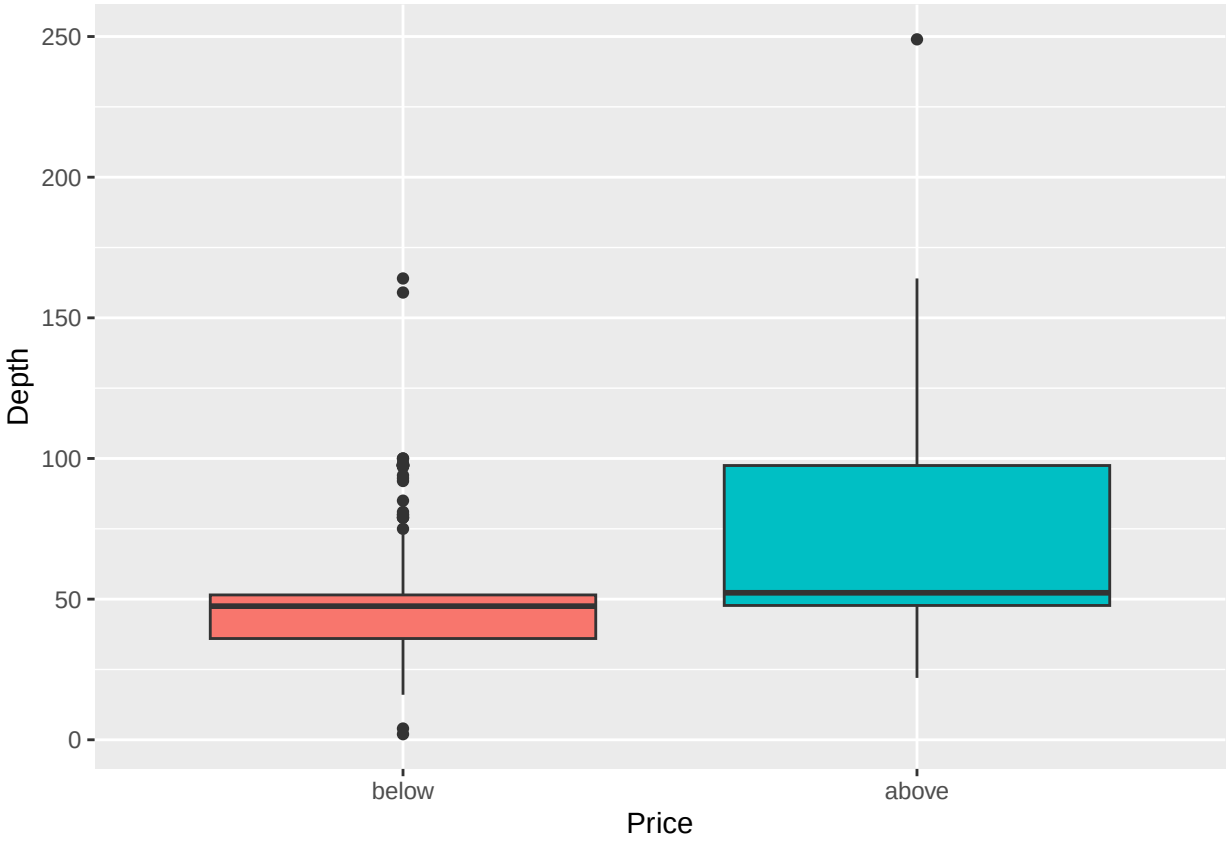
Relationship between other_colors and price.

```
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")
```



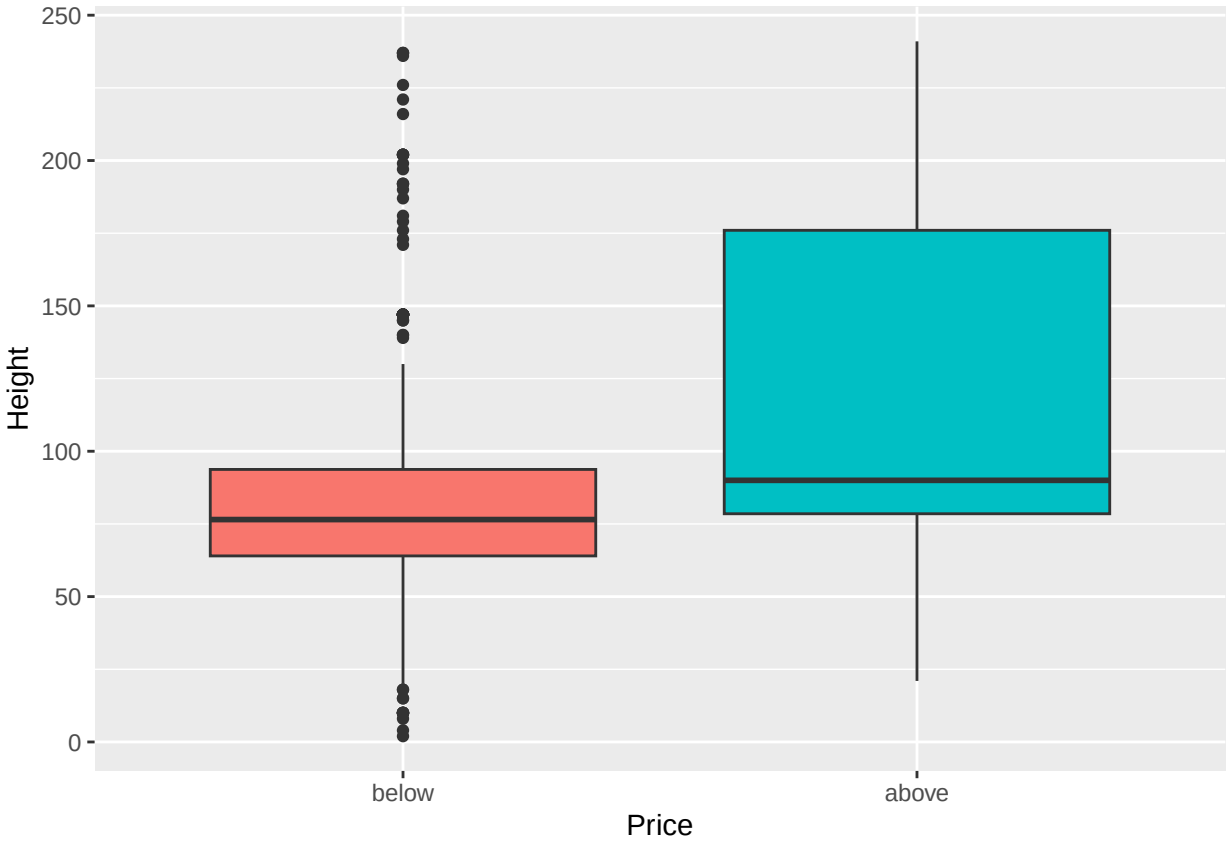
The proportion of above 1000 is higher for the other colors.

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

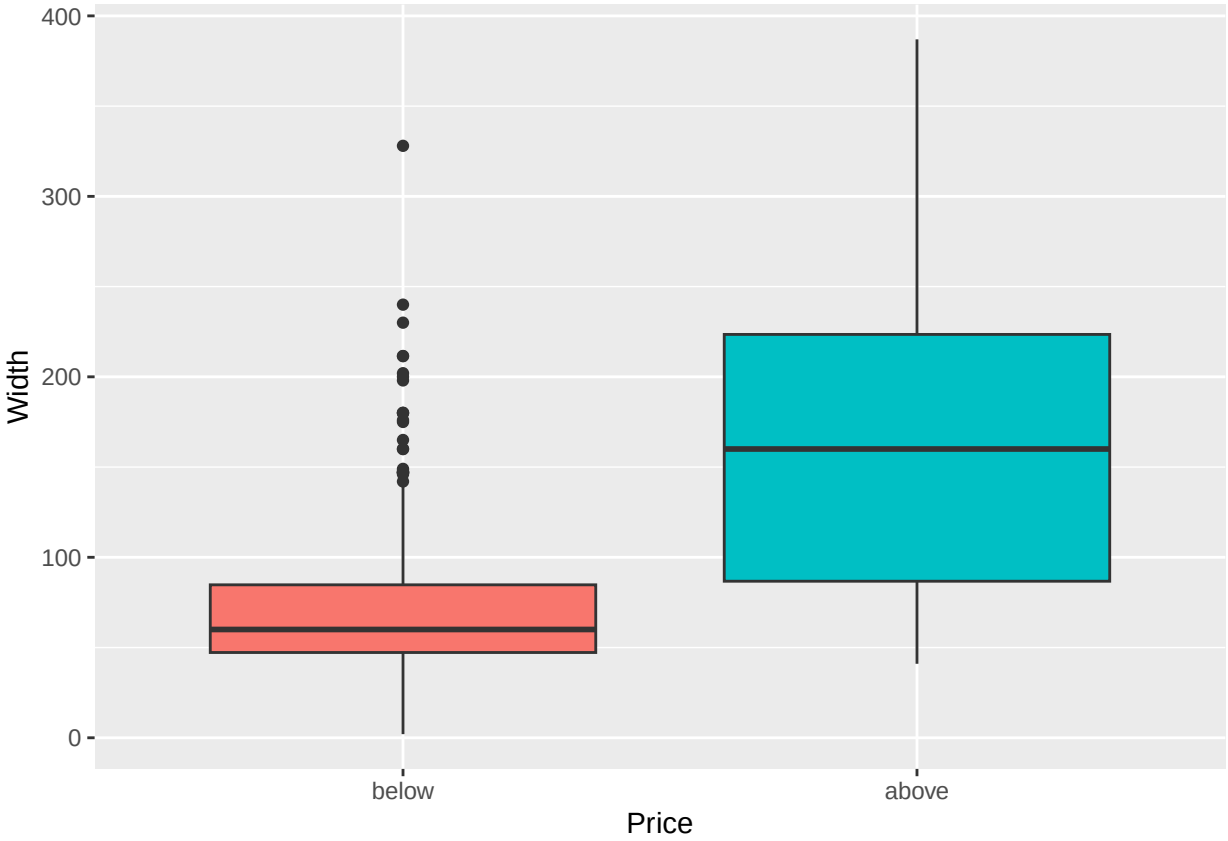


Depth between 50 and 100 seems like more possible to have price over1000

```
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 ~   sellable_online + other_colors + depth + height + width,
              data = furniture, family = binomial(link = "logit"))
summary(model1)
```

```
##
## Call:
## glm(formula = price ~ sellable_online + other_colors + depth +
##     height + width, family = binomial(link = "logit"), data = furniture)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4246  -0.6131  -0.4430   0.4717   2.1436
##
## Coefficients:
```

```
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -17.726645 626.020920  -0.028  0.97741
## sellable_onlineTRUE  13.056129 626.020876   0.021  0.98336
## other_colorsYes       0.008503   0.265881   0.032  0.97449
## depth                 0.022122   0.005490   4.030 5.59e-05 ***
## height                0.006967   0.002525   2.760  0.00579 **
## width                 0.018812   0.002550   7.379 1.60e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 587.56  on 457  degrees of freedom
## Residual deviance: 386.80  on 452  degrees of freedom
## AIC: 398.8
##
## Number of Fisher Scoring iterations: 14
```

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:

```
model2 <- glm(price~height+width+depth, furniture, family = "binomial")
summary(model2)
```

```
##
## Call:
## glm(formula = price ~ height + width + depth, family = "binomial",
##     data = furniture)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4340  -0.6108  -0.4438   0.4685   2.1522
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.699722   0.455735 -10.312  < 2e-16 ***
## height       0.007012   0.002526   2.776   0.0055 **
## width        0.018919   0.002537   7.459 8.74e-14 ***
## depth        0.022307   0.005442   4.099 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 587.56  on 457  degrees of freedom
## Residual deviance: 387.63  on 454  degrees of freedom
## AIC: 395.63
##
## Number of Fisher Scoring iterations: 5
```

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.

The intercept term (-4.699722) represents the estimated log-odds of the furniture price being more than 1000 Saudi Riyals when all the predictor variables are equal to zero (i.e., when the furniture has zero height, width, and depth). Since this intercept term is negative, it implies that the estimated probability of the furniture price being more than 1000 Saudi Riyals is low when the dimensions are zero.

The coefficient for the height variable (0.007012) suggests that, on average, a one-centimeter increase in height is associated with a 0.7% increase in the log-odds of the furniture price being more than 1000 Saudi Riyals, holding all other variables constant. This coefficient is positive, indicating that an increase in height is associated with a higher probability of the furniture price being more than 1000 Saudi Riyals.

The coefficient for the width variable (0.018919) implies that, on average, a one-centimeter increase in width is associated with a 1.9% increase in the log-odds of the furniture price being more than 1000 Saudi Riyals, holding all other variables constant. This coefficient is also positive, indicating that an increase in width is associated with a higher probability of the furniture price being more than 1000 Saudi Riyals.

The coefficient for the depth variable (0.022307) suggests that, on average, a one-centimeter increase in depth is associated with a 2.2% increase in the log-odds of the furniture price being more than 1000 Saudi Riyals, holding all other variables constant. This coefficient is also positive, indicating that an increase in depth is associated with a higher probability of the furniture price being more than 1000 Saudi Riyals.

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.