Group_21_Analysis

Group21

2023-03-13

1 Introduction

IKEA furniture is known for its modern and unusual designs, and the price of IKEA furniture is a significant concern for consumers when buying their furniture. The cost of IKEA furniture has a direct impact on consumer trust in the IKEA brand and IKEA's profits. The data in this report comes from official IKEA data, which documents the relationship between the price of different furniture products and other variables. Therefore, our team will study the relationship between the attributes of furniture and furniture with more than 1000 Saudi Riyals.

Section 2 contains a specific analysis of each variable in the data and discusses how to deal with missing values and investigate the co-linearity between variables. In section 3, we summarize the statistical values of the mean, minimum, etc., of each variable in the data and analyze the categorical factors and numerical variables using barplots and boxplots, respectively. Also, the relationship between categorical factors and numerical variables on our dependent variable, furniture price, was analyzed using barplot and boxplot, respectively. In section 4, we tested different models using AIC, Hoslem test, etc., until we finally selected the most appropriate model. In section 5, we summarize the relevant findings from the selected models, and we propose hypotheses and questions for the future work tasks in section 6.

1.1 Data description

Our data in this case is about furniture in IKEA Saudi Arabia.

The Features we picked are as below:

- category – The furniture category the item belongs to
- price – The current price in Saudi Riyals (as recorded on 20/04/2020)
- sellable_online – Is the item available to purchase online?
- other_colors – Is the item available in other colours
- depth – Depth of the item in cm
- height – Height of the item in cm
- width – width of the item in cm

2 Data preprocessing

Before analysing the data, we do the data preprocessing. In this part, we choose to use the variables' median to replace its missing value to get our final dataset.

```
#To see if there is colinearity in the numerical variables
ggpairs(group21_new,columns = c(4:6), title = "Correlation between numerical variables",)
```

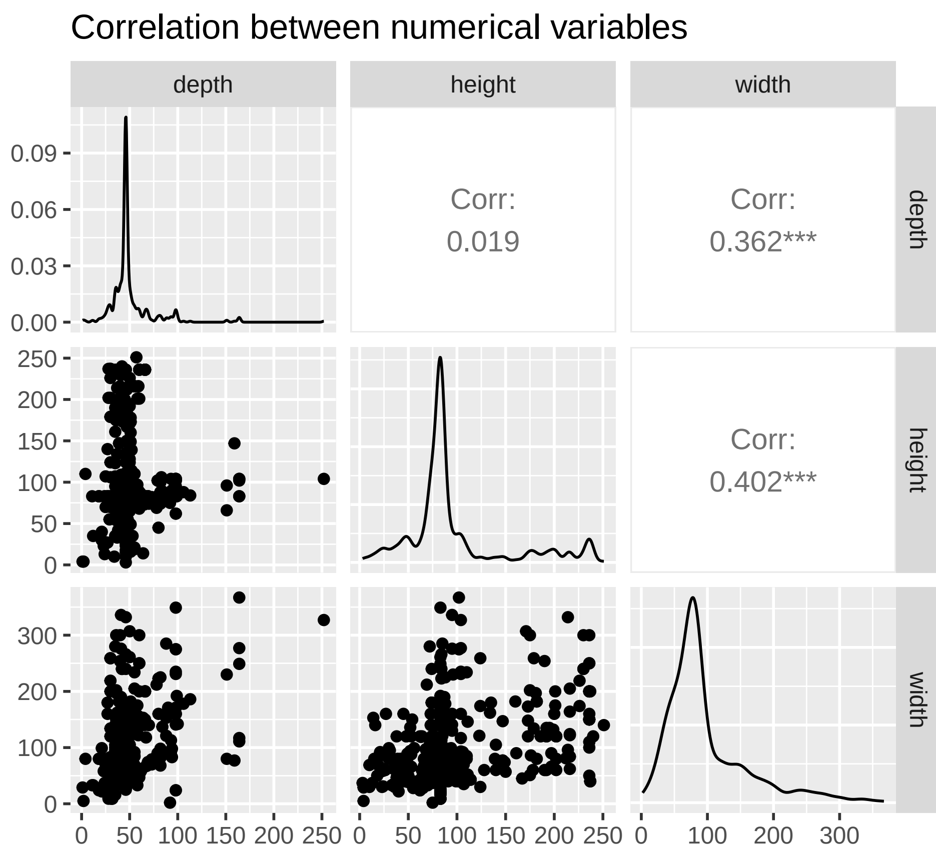

Figure 1: Correlation in numerical variables

Figure 1 shows that there's a weak relationship between `height` and `depth`(0.019), a mild relationship between `width` and `depth`(0.362) and a moderate relationship between `width` and `height`(0.402)

```
round(diag(var(group21_new[,4:6])),2) ##significant difference variance
```

```
##   depth  height   width
##  530.24 2746.02 3927.11
```

```
group21_new$depth <- scale(group21_new$depth,center=TRUE, scale=TRUE)
group21_new$height <- scale(group21_new$height,center=TRUE, scale=TRUE)
group21_new$width <- scale(group21_new$width,center=TRUE, scale=TRUE)
```

Based on the result above, we conclude that there is minimal variation between the `depth` of the different products in the data set, with more significant variation in the `height`. Still, the tremendous variation is in the `width` between the products. In other words, the table above means that the `depth` distribution is the smallest, followed by the `height`, and the `width` distribution is the largest. Finally, even the slightest `depth` variation has an enormous value, so we use the scale function below to normalize the data.

Then we use chi-square test to check if there is co-linearity in the three categorical variables

```
table_variable_1 <- table(group21_new$category, group21_new$sellable_online)
chisq_result_1 <- chisq.test(table_variable_1)
table_variable_2 <- table(group21_new$category, group21_new$other_colors)
chisq_result__2 <- chisq.test(table_variable_2)
table_variable_3 <- table(group21_new$other_colors, group21_new$sellable_online)
chisq_result_3 <- chisq.test(table_variable_3)
chisq_result_1 #p-value>0.05, there is no colinearity between category and sellable_online.
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_variable_1
## X-squared = 6.4756, df = 16, p-value = 0.9821
```

```
chisq_result__2#p-value<0.05,there is colinearity between category and other_colors.
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_variable_2
## X-squared = 101.66, df = 16, p-value = 1.693e-14
```

```
chisq_result_3#p-value>0.05, there is no colinearity between other_colors and sellable_online.
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_variable_3
## X-squared = 2.1624e-27, df = 1, p-value = 1
```

From the output above, co-llinearity only happens between `category` and `other_colors`, but not between other categorical variables.

3 Explanatory data Analysis

3.0.1 Summary statistics

```
#Summary statistic
summary(group21_new)
```

```
##                          category   sellable_online other_colors
##  Chairs                      : 74   FALSE:  1       No :300
##  Bookcases & shelving units: 71   TRUE :499       Yes:200
```

```
##  Tables & desks            : 67
##  Sofas & armchairs         : 51
##  Cabinets & cupboards      : 47
##  Wardrobes                 : 36
##  (Other)                   :154
##       depth.V1           height.V1            width.V1           price
##  Min.   :-2.151304   Min.   :-1.7867471   Min.   :-1.527478   Min.   :   3.0
##  1st Qu.:-0.370783   1st Qu.:-0.4318497   1st Qu.:-0.601947   1st Qu.: 168.8
##  Median :-0.197073   Median :-0.2601021   Median :-0.282798   Median : 457.0
##  Mean   : 0.000000   Mean   : 0.0000000   Mean   : 0.000000   Mean   : 991.1
##  3rd Qu.:-0.023364   3rd Qu.:-0.0072516   3rd Qu.: 0.355500   3rd Qu.:1245.0
##  Max.   : 8.748961   Max.   : 2.9458523   Max.   : 4.296987   Max.   :8551.0
##
##  newprice
##  more:354
##  less:146
##
##
##
##
##
```

The above summary of the four groups of variables shows that their maximum values differ significantly from the other data, so all four groups have outliers. And the central width portion is spread out furthest between three independent variables.

3.0.2 Boxplot

A Boxplot is used here to find the relationship between our response variable `newprice` and numeric variables.

```r
#Boxplot
box1<-ggplot(data = group21_new, aes(x = newprice, y = depth , fill = newprice))+
  geom_boxplot() +
  labs(x = "More or less", y = "Depth")+
  theme(legend.position = "none")

box2<-ggplot(data = group21_new, aes(x = newprice, y = height , fill = newprice))+
  geom_boxplot() +
  labs(x = "More or less", y = "Height")+
  theme(legend.position = "none")

box3<-ggplot(data = group21_new, aes(x = newprice, y = width , fill = newprice))+
  geom_boxplot() +
  labs(x = "More or less", y = "Width")+
  theme(legend.position = "none")
grid.arrange(box1,box2,box3,ncol=3)
```

Figure 2 shows three different variables (`depth`, `height`, `width`) in `newprice`. It can be said that there is more difference in Plot 3 between `newprice` and `width`, meaning the `width` will influence the price more than the other two variables.

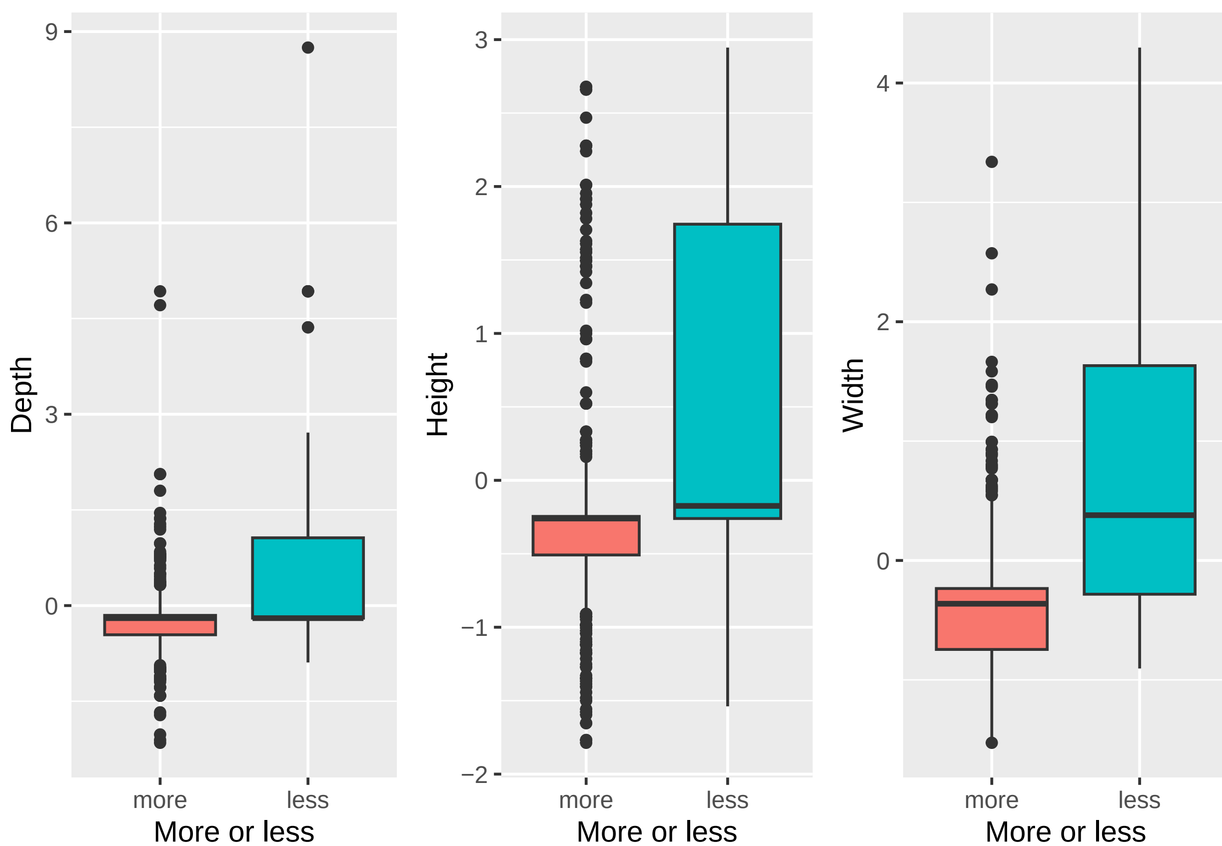

Figure 2: Boxplot of newprice by numeric vairalbes

3.0.3 Barplot

A barplot is here used to determine the connection between categorical factors and the newprice.

```
#the proportion and barplots(categorical variables)
group21_new %>%
  tabyl(newprice, category) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns()
```

```
##  newprice Bar furniture      Beds Bookcases & shelving units
##      more      1.7% (6) 4.8% (17)                 17.5% (62)
##      less      1.4% (2) 8.9% (13)                  6.2%  (9)
##  Cabinets & cupboards Caf<e9> furniture     Chairs
##            10.5% (37)          0.8% (3) 15.3% (54)
##             6.8% (10)          0.0% (0) 13.7% (20)
##  Chests of drawers & drawer units Children's furniture Nursery furniture
##                         3.7% (13)            4.0% (14)         5.1% (18)
##                         0.0%  (0)            0.0%  (0)         0.0%  (0)
##  Outdoor furniture Room dividers Sideboards, buffets & console tables
##          6.5% (23)      0.0% (0)                             0.6% (2)
##          6.8% (10)      0.7% (1)                             1.4% (2)
##  Sofas & armchairs Tables & desks Trolleys TV & media furniture  Wardrobes
##          7.1% (25)     13.0% (46) 0.6% (2)            5.1% (18)  4.0% (14)
##         17.8% (26)     14.4% (21) 0.7% (1)            6.2%  (9) 15.1% (22)
```

```
group21_new %>%
  tabyl(sellable_online, newprice) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns() # To show original counts
```

```
##  sellable_online        more         less
##            FALSE  0.0%   (0) 100.0%   (1)
##             TRUE 70.9% (354)  29.1% (145)
```

```
group21_new %>%
  tabyl(other_colors, newprice) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns() # To show original counts
```

```
##  other_colors        more       less
##            No 75.7% (227) 24.3% (73)
##           Yes 63.5% (127) 36.5% (73)
```

```
#Barplot
bar1<-ggplot()+
  geom_bar(data =group21_new,
           aes(x = factor(category),fill = factor(newprice)),
           position = "fill")+
```

```
  labs(x = "category", y = "newprice")
bar1
```

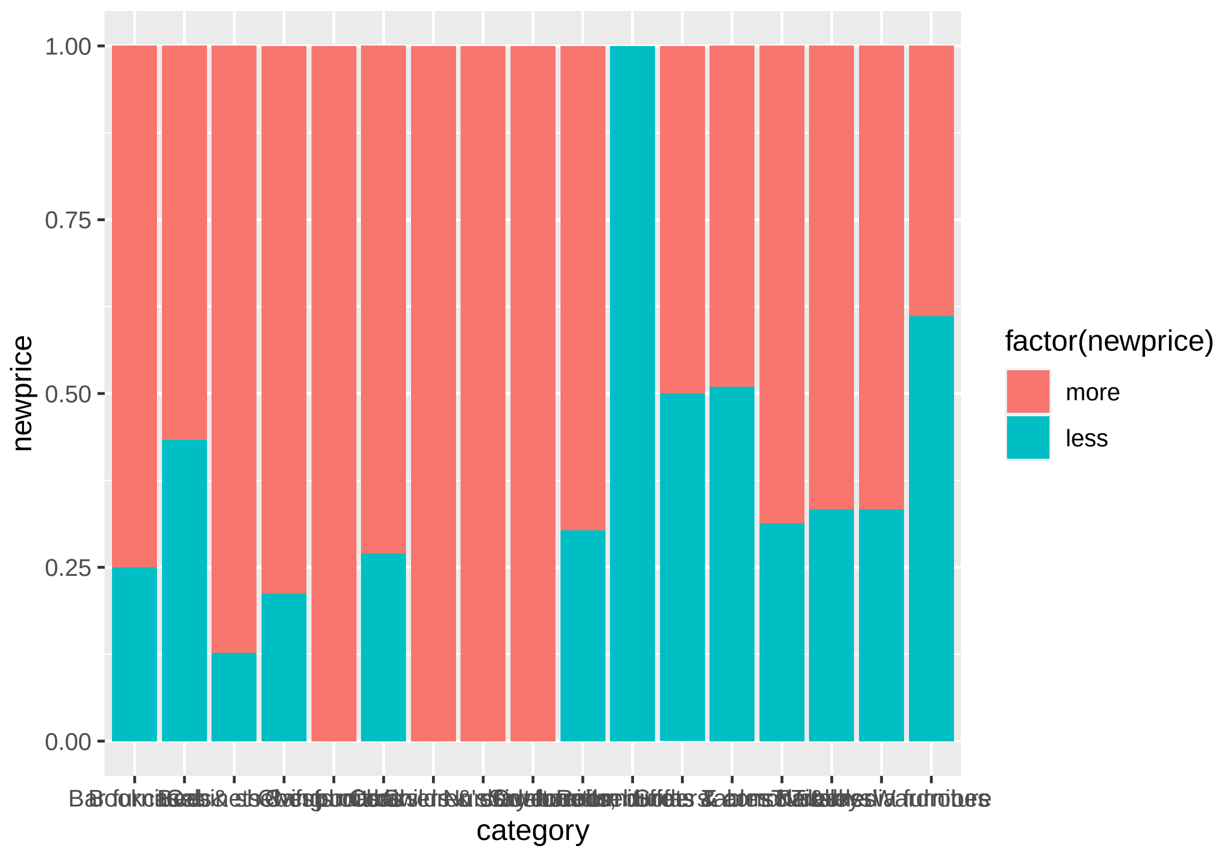

Figure 3: Boxplot of newprice by categorical vairalbes(category)

Figure 3, for all categories of furniture, the proportion of most of the “less” is less than 0.5, indicating that the ratio of more than 1000 Saudi Riyals exceeds the balance of less than 1000 Saudi Riyals in almost all furniture, so there is no apparent connection between `category` and our dependent variable `new_price`, and for our analysis, `category` is not a representative variable of whether there is an association between `category` and greater than 1000 Saudi Riyals.

```
bar2<-ggplot()+
  geom_bar(data=group21_new,
           aes(x=factor(sellable_online),fill = factor(newprice)),
           position="fill")+
  labs(x = "sellable_online", y = "newprice")
bar2
```

Figure 4,in the “false” part of the above chart, there are no furniture items larger than 1000 Saudi Riyals, which means that all furniture items not sold online are smaller than 1000 Saudi Riyals. Therefore, using `sellable_online` as the independent variable cannot be entirely explained by the relationship with `new_price`.

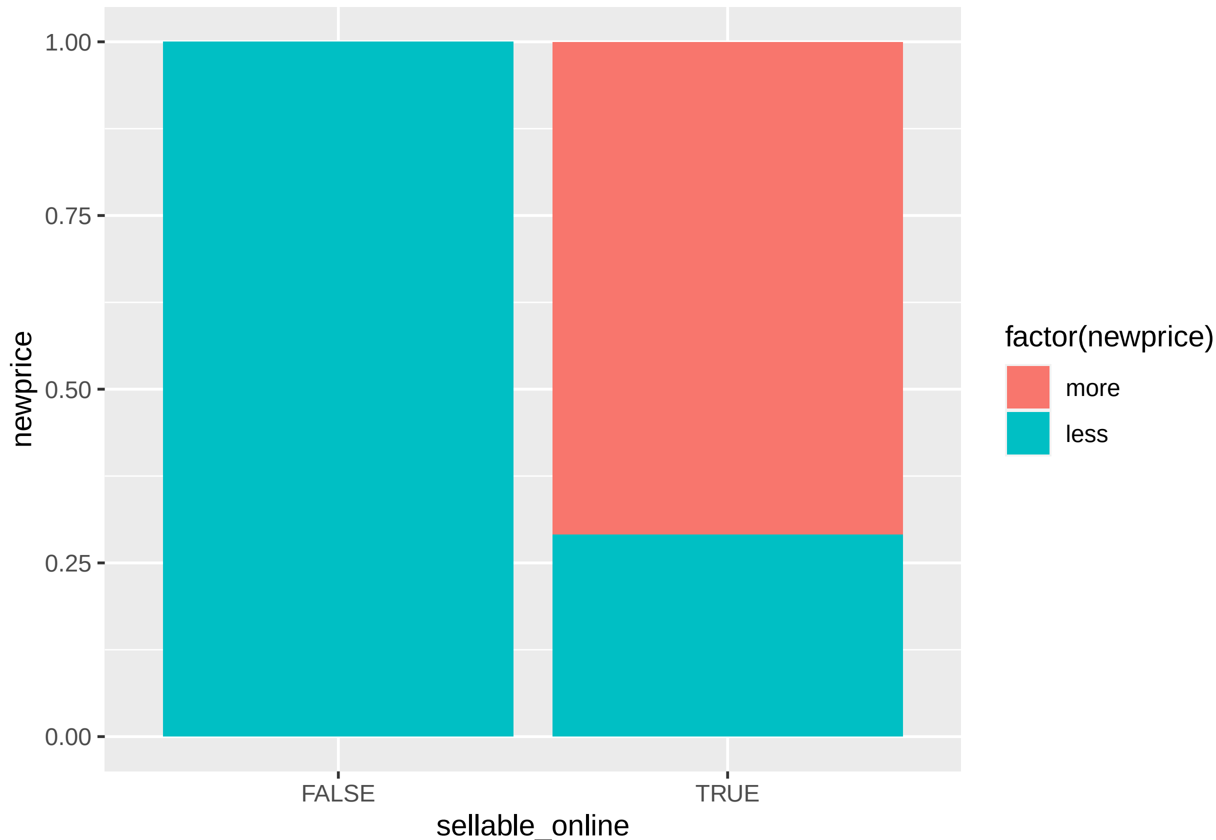

Figure 4: Boxplot of newprice by categorical vairalbes(sellable_online)

```
bar3<-ggplot()+
  geom_bar(data =group21_new,
           aes(x = factor(other_colors),fill = factor(newprice)),
           position = "fill")+
  labs(x = "sellable_online", y = "newprice")

bar3
```

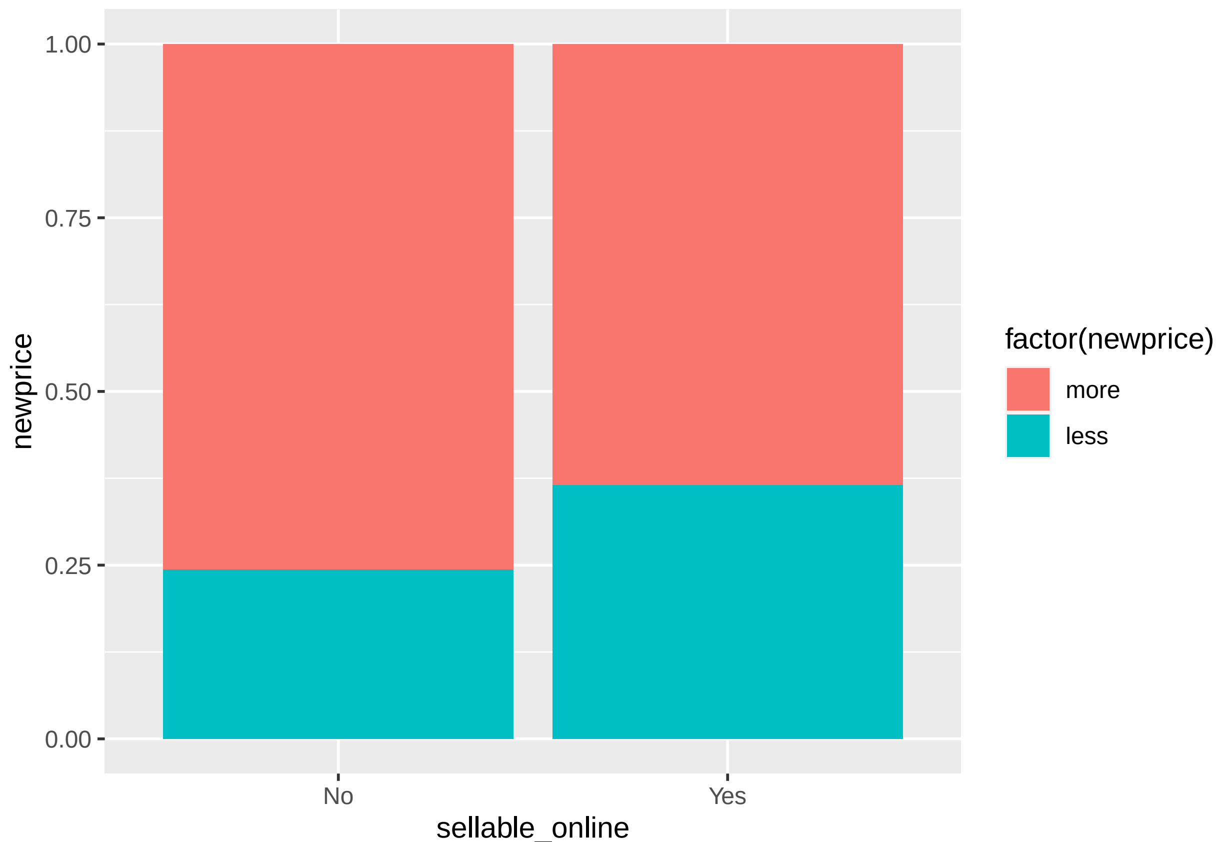

Figure 5: Boxplot of newprice by categorical vairalbes(other_colors)

Figure 5, finally, the table above shows that `other_colors` have a certain percentage of Saudi Riyals, whether larger than 1000 or not, and the difference between the two groups is insignificant. Therefore, we can use `other_colors` as the central modelling premise.

4 Formal data analysis

In this part we first use the model selection method to pick the our best fitted model from the full model:

```
#Fit a GLM with all variables
model_1 <- glm(newprice ~ category+sellable_online+depth+height+width, data = group21_new,
               family = binomial(link = "logit"))

#Use stepwise selection to select the most important variables
step_model_1 <- step(model_1,direction = "both")#model selection(AIC value)
```

```
## Start:  AIC=366.84
## newprice ~ category + sellable_online + depth + height + width
##
##                   Df Deviance    AIC
## - depth            1   326.70 366.70
## - sellable_online  1   326.77 366.77
## <none>                 324.84 366.84
## - height           1   357.47 397.47
## - category        16   424.40 434.40
## - width            1   437.04 477.04
##
## Step:  AIC=366.7
## newprice ~ category + sellable_online + height + width
##
##                   Df Deviance    AIC
## - sellable_online  1   328.61 366.61
## <none>                 326.70 366.70
## + depth            1   324.84 366.84
## - height           1   361.08 399.08
## - category        16   448.72 456.72
## - width            1   461.56 499.56
##
## Step:  AIC=366.61
## newprice ~ category + height + width
##
##                   Df Deviance    AIC
## <none>                 328.61 366.61
## + sellable_online  1   326.70 366.70
## + depth            1   326.77 366.77
## - height           1   363.28 399.28
## - category        16   452.10 458.10
## - width            1   463.56 499.56
```

Based on model selection above, the model `newprice~category+height+width` with smallest AIC=366.61 is the best fitted model.

```
#Based on AIC pick the best model
model_2 <- glm(newprice ~ category+height+width, data = group21_new,
               family = binomial(link = "logit"))

#HLtest
hl_1 <- hoslem.test(group21_new$newprice, fitted(model_2), g=10)
hl_1 #p=2.313e-05<0.05, model_2 is not fitted well.
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  group21_new$newprice, fitted(model_2)
## X-squared = 35.345, df = 8, p-value = 2.313e-05
```

By checking the p-value from HL test, we can know that this model doesn't fit well.

Since `category` and `other_colors` are not independent, so next part we remove the `category` variable from the full model and use AIC to check it again.

```r
model_3 <- glm(newprice ~ sellable_online+other_colors+depth+height+width,
               data = group21_new, family = binomial(link = "logit"))

step_model_3 <- step(model_3,direction = "both")
```

```
## Start:  AIC=436.26
## newprice ~ sellable_online + other_colors + depth + height +
##     width
##
##                   Df Deviance    AIC
## - other_colors     1   424.40 434.40
## <none>                 424.26 436.26
## - sellable_online  1   427.92 437.92
## - height           1   432.62 442.62
## - depth            1   447.59 457.59
## - width            1   490.63 500.63
##
## Step:  AIC=434.4
## newprice ~ sellable_online + depth + height + width
##
##                   Df Deviance    AIC
## <none>                 424.40 434.40
## - sellable_online  1   428.01 436.01
## + other_colors     1   424.26 436.26
## - height           1   432.64 440.64
## - depth            1   448.72 456.72
## - width            1   492.96 500.96
```

```r
model_4 <- glm(newprice ~ sellable_online+depth+height+width, data = group21_new,
               family = binomial(link = "logit"))

#HLtest
hl_2 <- hoslem.test(group21_new$newprice, fitted(model_4), g=10)
hl_2  #p=0.000146<0.05, model_4 is not fitted well.
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  group21_new$newprice, fitted(model_4)
## X-squared = 30.907, df = 8, p-value = 0.000146
```

From the smallest AIC(434.4), we can tell that `other_color` won't influence the model, so we could remove it from the model. However, from the p-value of the HL test, this model doesn't fit well as well.

Then we want to increase the complexity to our model.

```r
model_5 <- glm(newprice ~ category+sellable_online+depth+height+width+I(depth^2)+I(depth^3),
               data = group21_new, family = binomial(link = "logit"))



step_model_4<- step(model_5,direction = "both")
```

```
## Start:  AIC=369.07
## newprice ~ category + sellable_online + depth + height + width +
##     I(depth^2) + I(depth^3)
##
##                   Df Deviance    AIC
## - depth            1   323.13 367.13
## - I(depth^3)       1   324.47 368.47
## - I(depth^2)       1   324.83 368.83
## - sellable_online  1   325.00 369.00
## <none>                 323.07 369.07
## - height           1   356.71 400.71
## - category        16   417.82 431.82
## - width            1   435.21 479.21
##
## Step:  AIC=367.13
## newprice ~ category + sellable_online + height + width + I(depth^2) +
##     I(depth^3)
##
##                   Df Deviance    AIC
## - I(depth^3)       1   324.57 366.57
## - sellable_online  1   325.05 367.05
## <none>                 323.13 367.13
## - I(depth^2)       1   325.70 367.70
## + depth            1   323.07 369.07
## - height           1   357.12 399.12
## - category        16   443.07 455.07
## - width            1   446.67 488.67
##
## Step:  AIC=366.57
## newprice ~ category + sellable_online + height + width + I(depth^2)
##
##                   Df Deviance    AIC
## - sellable_online  1   326.49 366.49
## <none>                 324.57 366.57
## - I(depth^2)       1   326.70 366.70
## + I(depth^3)       1   323.13 367.13
## + depth            1   324.47 368.47
## - height           1   358.46 398.46
## - category        16   443.73 453.73
## - width            1   449.99 489.99
##
## Step:  AIC=366.49
## newprice ~ category + height + width + I(depth^2)
##
##                   Df Deviance    AIC
## <none>                 326.49 366.49
## + sellable_online  1   324.57 366.57
## - I(depth^2)       1   328.61 366.61
## + I(depth^3)       1   325.05 367.05
## + depth            1   326.40 368.40
## - height           1   360.67 398.67
## - category        16   447.22 455.22
## - width            1   452.01 490.01
```

```r
model_6 <- glm(newprice ~ category + height + width + I(depth^2), data = group21_new,
               family = binomial(link = "logit"))

#HLtest
hl_3 <- hoslem.test(group21_new$newprice, fitted(model_6), g=10)
hl_3  #p=0.0005213<0.05, model_6 is not fitted well.
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  group21_new$newprice, fitted(model_6)
## X-squared = 27.764, df = 8, p-value = 0.0005213
```

```r
model_7 <- glm(newprice ~ category+sellable_online+depth+height+width+I(height^2)
               +I(height^3), data = group21_new, family = binomial(link = "logit"))

step_model_5<- step(model_7,direction = "both")
```

```
## Start:  AIC=368.36
## newprice ~ category + sellable_online + depth + height + width +
##     I(height^2) + I(height^3)
##
##                   Df Deviance    AIC
## - I(height^2)      1   322.53 366.53
## - I(height^3)      1   322.90 366.90
## <none>                 322.36 368.36
## - sellable_online  1   324.38 368.38
## - depth            1   324.64 368.64
## - height           1   326.38 370.38
## - category        16   413.37 427.37
## - width            1   435.11 479.11
##
## Step:  AIC=366.53
## newprice ~ category + sellable_online + depth + height + width +
##     I(height^3)
##
##                   Df Deviance    AIC
## - sellable_online  1   324.52 366.52
## <none>                 322.53 366.53
## - depth            1   324.76 366.76
## - I(height^3)      1   324.84 366.84
## + I(height^2)      1   322.36 368.36
## - height           1   326.38 368.38
## - category        16   419.24 431.24
## - width            1   435.70 477.70
##
## Step:  AIC=366.52
## newprice ~ category + depth + height + width + I(height^3)
##
##                   Df Deviance    AIC
## <none>                 324.52 366.52
## + sellable_online  1   322.53 366.53
```

```
## - depth               1   326.72 366.72
## - I(height^3)         1   326.77 366.77
## + I(height^2)         1   324.38 368.38
## - height              1   328.51 368.51
## - category           16   422.89 432.89
## - width               1   437.82 477.82
```

```r
model_8 <- glm(newprice ~ category + depth + height + width + I(height^3),
               data = group21_new, family = binomial(link = "logit"))

#HLtest
hl_4 <- hoslem.test(group21_new$newprice, fitted(model_8), g=10)
hl_4  #p=0.001637<0.05, model_8 is not fitted well.
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  group21_new$newprice, fitted(model_8)
## X-squared = 24.868, df = 8, p-value = 0.001637
```

```r
##wide^3
model_9<- glm(newprice ~ category+sellable_online+depth+height+width+I(width^2)+I(width^3),
              data = group21_new, family = binomial(link = "logit"))

step_model_6<- step(model_9,direction = "both")
```

```
## Start:  AIC=362.19
## newprice ~ category + sellable_online + depth + height + width +
##     I(width^2) + I(width^3)
##
##                   Df Deviance    AIC
## - depth            1   317.73 361.73
## - sellable_online  1   318.01 362.01
## <none>                 316.19 362.19
## - I(width^3)       1   320.59 364.59
## - I(width^2)       1   323.81 367.81
## - height           1   346.33 390.33
## - width            1   376.23 420.23
## - category        16   409.72 423.72
##
## Step:  AIC=361.73
## newprice ~ category + sellable_online + height + width + I(width^2) +
##     I(width^3)
##
##                   Df Deviance    AIC
## - sellable_online  1   319.53 361.53
## <none>                 317.73 361.73
## + depth            1   316.19 362.19
## - I(width^3)       1   321.83 363.83
## - I(width^2)       1   325.31 367.31
## - height           1   349.16 391.16
## - width            1   390.63 432.63
## - category        16   433.42 445.42
```

```
##
## Step:  AIC=361.53
## newprice ~ category + height + width + I(width^2) + I(width^3)
##
##                    Df Deviance    AIC
## <none>                   319.53 361.53
## + sellable_online  1   317.73 361.73
## + depth            1   318.01 362.01
## - I(width^3)       1   323.70 363.70
## - I(width^2)       1   327.22 367.22
## - height           1   351.21 391.21
## - width            1   392.46 432.46
## - category        16   436.50 446.50
```

```
model_10 <- glm(newprice ~ category + height + width + I(width^2) + I(width^3),
                data = group21_new, family = binomial(link = "logit"))

#HLtest
hl_5 <- hoslem.test(group21_new$newprice, fitted(model_10), g=10)
hl_5  #p=0.336>0.05, model_10 is fitted well.
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  group21_new$newprice, fitted(model_10)
## X-squared = 9.0753, df = 8, p-value = 0.336
```

```
#wide^4
model_11<- glm(newprice ~ category+sellable_online+depth+height+width+I(width^2)
               +I(width^3)+I(width^4), data = group21_new, family = binomial(link = "logit"))

step_model_7<- step(model_11,direction = "both")
```

```
## Start:  AIC=354.47
## newprice ~ category + sellable_online + depth + height + width +
##     I(width^2) + I(width^3) + I(width^4)
##
##                    Df Deviance    AIC
## - sellable_online  1   308.19 354.19
## <none>                   306.47 354.47
## - depth            1   309.05 355.05
## - I(width^4)       1   316.19 362.19
## - I(width^3)       1   319.83 365.83
## - I(width^2)       1   321.44 367.44
## - width            1   332.84 378.84
## - height           1   337.87 383.87
## - category        16   401.08 417.08
##
## Step:  AIC=354.19
## newprice ~ category + depth + height + width + I(width^2) + I(width^3) +
##     I(width^4)
##
##                    Df Deviance    AIC
```

```
## <none>                   308.19 354.19
## + sellable_online  1   306.47 354.47
## - depth             1   310.75 354.75
## - I(width^4)        1   318.01 362.01
## - I(width^3)        1   321.71 365.71
## - I(width^2)        1   323.35 367.35
## - width             1   334.50 378.50
## - height            1   339.83 383.83
## - category         16   404.13 418.13
```

```
model_12 <- glm(newprice ~ category + depth + height + width + I(width^2) + I(width^3)
                +I(width^4), data = group21_new, family = binomial(link = "logit"))

#HLtest
hl_6 <- hoslem.test(group21_new$newprice, fitted(model_12), g=10)
hl_6  #p=0.5247>0.05,model_12 is fitted well
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  group21_new$newprice, fitted(model_12)
## X-squared = 7.1115, df = 8, p-value = 0.5247
```

```
#wide^4,depth^2
model_13<- glm(newprice ~ category+sellable_online+depth+height+width+I(width^2)
               +I(width^3)+I(width^4)+I(depth^2),
               data = group21_new, family = binomial(link = "logit"))

step_model_8<- step(model_13,direction = "both")
```

```
## Start:  AIC=356.18
## newprice ~ category + sellable_online + depth + height + width +
##     I(width^2) + I(width^3) + I(width^4) + I(depth^2)
##
##                   Df Deviance    AIC
## - depth            1   306.42 354.42
## - I(depth^2)       1   306.47 354.47
## - sellable_online  1   307.89 355.89
## <none>                 306.18 356.18
## - I(width^4)       1   315.82 363.82
## - I(width^3)       1   319.30 367.30
## - I(width^2)       1   320.92 368.92
## - width            1   332.84 380.84
## - height           1   337.83 385.83
## - category        16   395.10 413.10
##
## Step:  AIC=354.42
## newprice ~ category + sellable_online + height + width + I(width^2) +
##     I(width^3) + I(width^4) + I(depth^2)
##
##                   Df Deviance    AIC
## - sellable_online  1   308.12 354.12
## <none>                 306.42 354.42
```

```
## - I(depth^2)       1   309.05 355.05
## + depth            1   306.18 356.18
## - I(width^4)       1   315.87 361.87
## - I(width^3)       1   319.30 365.30
## - I(width^2)       1   321.01 367.01
## - width            1   335.97 381.97
## - height           1   338.82 384.82
## - category        16   420.82 436.82
##
## Step:  AIC=354.12
## newprice ~ category + height + width + I(width^2) + I(width^3) +
##     I(width^4) + I(depth^2)
##
##                   Df Deviance    AIC
## <none>                 308.12 354.12
## + sellable_online  1   306.42 354.42
## - I(depth^2)       1   310.75 354.75
## + depth            1   307.89 355.89
## - I(width^4)       1   317.68 361.68
## - I(width^3)       1   321.17 365.17
## - I(width^2)       1   322.90 366.90
## - width            1   337.59 381.59
## - height           1   340.77 384.77
## - category        16   423.80 437.80
```

```r
model_14 <- glm(newprice ~ category + height + width+I(width^2)+I(width^3)+I(width^4)
                +I(depth^2), data = group21_new, family = binomial(link = "logit"))

#HLtest
hl_7 <- hoslem.test(group21_new$newprice, fitted(model_14), g=10)
hl_7  #p=0.527>0.05, model_14 is fitted well.
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  group21_new$newprice, fitted(model_14)
## X-squared = 7.0898, df = 8, p-value = 0.527
```

4.1 Any thoughts on HL test

Based on our knowledge, the Hosmer-Lemeshow test is used for testing model goodness of fit and the test is used in the chi-square test with g-2 degrees of freedom. By running the Hosmer-Lemeshow test, a large chi-squared value with a p-value less than 0.05 indicates a poor fit and p- a value close to 1 indicates an excellent logistic regression model fit. When we first generalized all the independent variables in the dataset to the dependent variable new_price, we found that the p-value was too small and less than 0.05. So we kept trying to extract a few of the independent variables to model new_price, and finally, all the p-values were less than 0.05 and in a small range. The above steps proves that more than simple power-of-one modelling of the independent variables is needed to build a suitable model. Then we add new variables to `depth`, `width` and `height`, respectively, and we find that increasing the power of `depth` and `height` is still less than 0.05 for modelling the p-value, but increasing the capacity of width keeps the p-value close to 1, which means we can make our model fit better. In summary, increasing the `width` strength is the most critical variable in our modelling based on the dataset.

5 Conclusions

When increasing the complexity of the model, it is most beneficial to increase the complexity of the `width` to improve the model's fit directly. The final model build tells us that the `width` variable is an important variable that affects the dependent variable `new_price` and that the p-value of the model becomes more significant as the exponential value of width is increased.

For IKEA furniture larger than 1000 Saudi Riyals, `category` and `sellable_online` have a lot of bias for analyzing the model. In contrast, `other_colors` can help build the most comprehensive relationship with the dependent variable and make our model more convincing.

6 Further extension

For the further extension, we could check the GLMs' model assumption , to see if the model really fits well. Also could use other method to detect all the models that seems to fit well from the formal analysis, to decide which one will perform the best from those models.