Determining the Influence of Different Variables on the Price of Ikea Products – a Regression Analysis

Philip Krück, Johannes Pein

Hamburg School of Business Administration

B.Sc. Business Informatics (A Track)

Digital Toolbox: Data Business

Lecturer: Ulf Köther

Group Number: 7

*Matriculation Numbers: 3938 (P.Krück), **** (J.Pein)*

04.12.2020

Contents

Appendices

List of Abbreviations

R	Statistical Programming Language.

1
Introduction

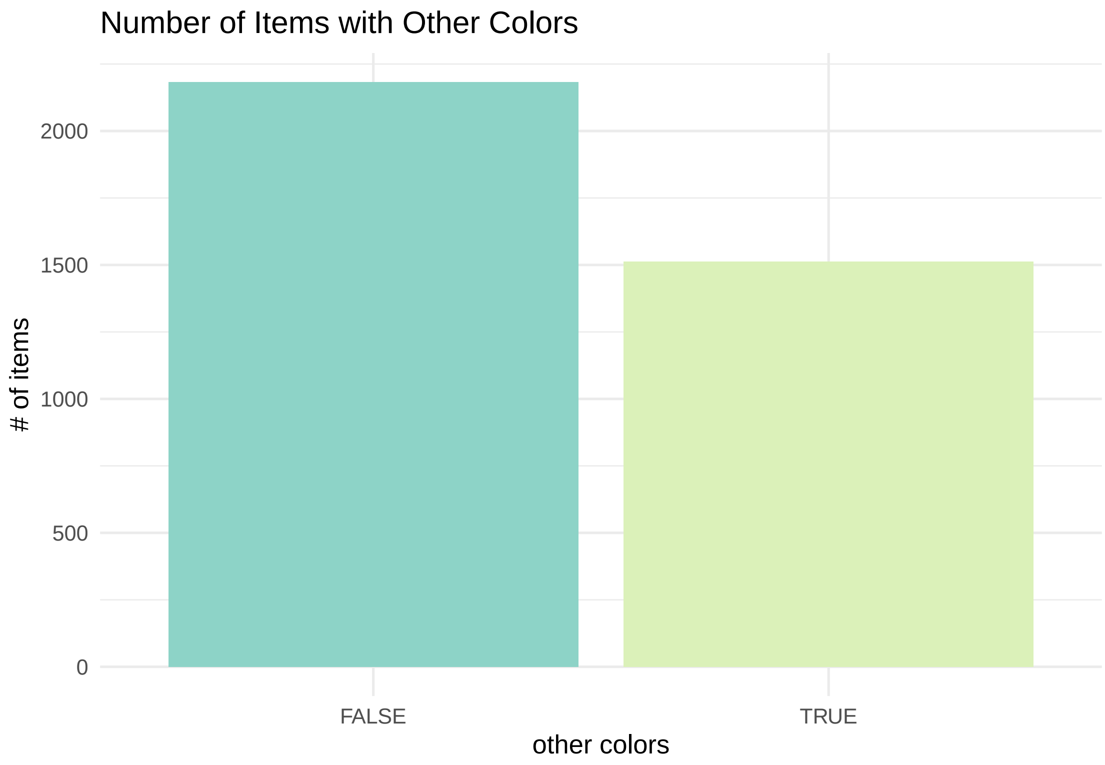

Figure 1.1: Sample plot

2
Theoretical Background & Research Question

2.1 Theoretical Background

2.1.1 Data Set

The data set was obtained by a kaggle.com user (Reem Abdulrahman) by the means of webscraping techniques from the Saudi Arabian Ikea website in the furniture category on the 20th of April 2020. Noteworthy features include the name, category, price in Saudi Riyals, the designer and dimensions (width, height and depth). The data set has 13 variables and 2962 observations.

2.1.2 Overfitting

->Johannes

2.1.3 RF Basics

->Johannes

2.1.4 Feature Importance

->Johannes

2.2 Research Question

This paper explores the influence for different variables on the price in the given data set. The motivating forces for this research question are the possible implications for price determination of new items.

3
Methods

3.1 Data Cleaning and Transformatoin

To examine the given data set properly, the authors first had to restructure and reformat it. This initial data cleaning step included type conversion, value mutation, addition of newly calculated fields and the removal of irrelevant columns. Concretely, name, category and designer were converted to categorical variables. In the designer column, blank strings and values prefixed by "IKEA of Sweden" were converted to missing values (`NA`). Furthermore, both the price and old price were converted to double values and the currency was changed from Saudi Arabian Riyals to Euros based on the exchange rate from the time the data set was obtained by the author @ref(#theoretical_background).

Interestingly, the data set had a peculiarity where some rows were exact duplicates except for the category value. The authors considered multiple approaches to handle these data duplications without losing information about the category of an item.

One considered option was to merge the two category values into one column value via comma separation (e.g. `"a"` and `"b"` converts to `"a, b"`). However, this approach leads to the creation of many combinatorial categories with a low count of items per category which also reduces the item count per category where

Table 3.1: Initial Data Set formatting.

X1	item_id	name	category	price	old_price	sellabl
0	90420332	FREKVENS	Bar furniture	2650	No old price	TRUE
1	368814	NORDVIKEN	Bar furniture	9950	No old price	FALSE
2	9333523	NORDVIKEN / NORDVIKEN	Bar furniture	20950	No old price	FALSE
3	80155205	STIG	Bar furniture	690	No old price	TRUE
4	30180504	NORBERG	Bar furniture	2250	No old price	TRUE
5	10122647	INGOLF	Bar furniture	3450	No old price	TRUE

the category isn't comma separated. Overall this would lead to having many small categories which increases the difficulty in applying a regression model due to overfitting @ref(#overfitting).

The second option was to create separate columns for the different values of `category`. The data set would then have observations with category one, two and three. While no information is lost utilizing this approach, most observations in the second and third category column would contain missing values, thus increasing the difficulty of analysis using a predefined model @**??**#random_forest_model).

The authors chose the option of selecting the observations out of the duplicates where the category count occurred most frequently when considering duplicates. The most important categories could be retained without including more column vectors into the data set as in option two.

To better facilitate the comparison of the different sizes of furniture items, the size in cubic meters was computed based on the depth, width and height values, and added as a column vector for further analysis. Finally, the authors only selected columns that could have a potential impact on the analysis @ref(#research_question) for further investigation. A detailed comparison of the initial vs. transformed data structure can be seen in tables 3.1 and 3.2.

TODO: Format these tables

Table 3.2: Data Set after cleaning process.

name	category	price_eur	old_price_eur	sellable_online
FREKVENS	Bar furniture	65.02	NA	TRUE
NORDVIKEN	Bar furniture	244.14	NA	FALSE
NORDVIKEN / NORDVIKEN	Bar furniture	514.05	NA	FALSE
STIG	Bar furniture	16.93	NA	TRUE
NORBERG	Bar furniture	55.21	NA	TRUE
INGOLF	Bar furniture	84.65	NA	TRUE

3.2 Exploratory Data Analysis

The following sections explore our data based on the eight step data exploration protocol proposed by Zuur et al (lain Zuur 2010).

3.2.1 Step 1: Outliers in Price and Independent Variables

Outliers of the chosen variables (from tidying see @ref(#datacleaning) can be observed for each variable (see plot . . .). The authors assume that outliers do not occur randomly in the form of an observer error. Web scraping code is written in a generic form which makes it generalizable to all applied pages. Thus takes human observation errors out of the equation. Additionally, the authors looked at individual outlier (stichprobenartig) examples and used the provided link column to manually double check observations against machine errors. By the stated assumption, all outliers are meaningful for further analysis.

3.2.2 Step 2: Homogeneity of Price

The homogeneity (homoscedasticity) of variance for price is explored by the means of conditional boxplotting. Within each name, and within each category the variance is heterogenous (3.1). However, looking at both name and category in conjunction, it is possible to explore homoscedasticity of variance for price. In the context of this paper, the authors weren't able to inspect all variable combinations for the five categorical variables ($2^3 = 8$).

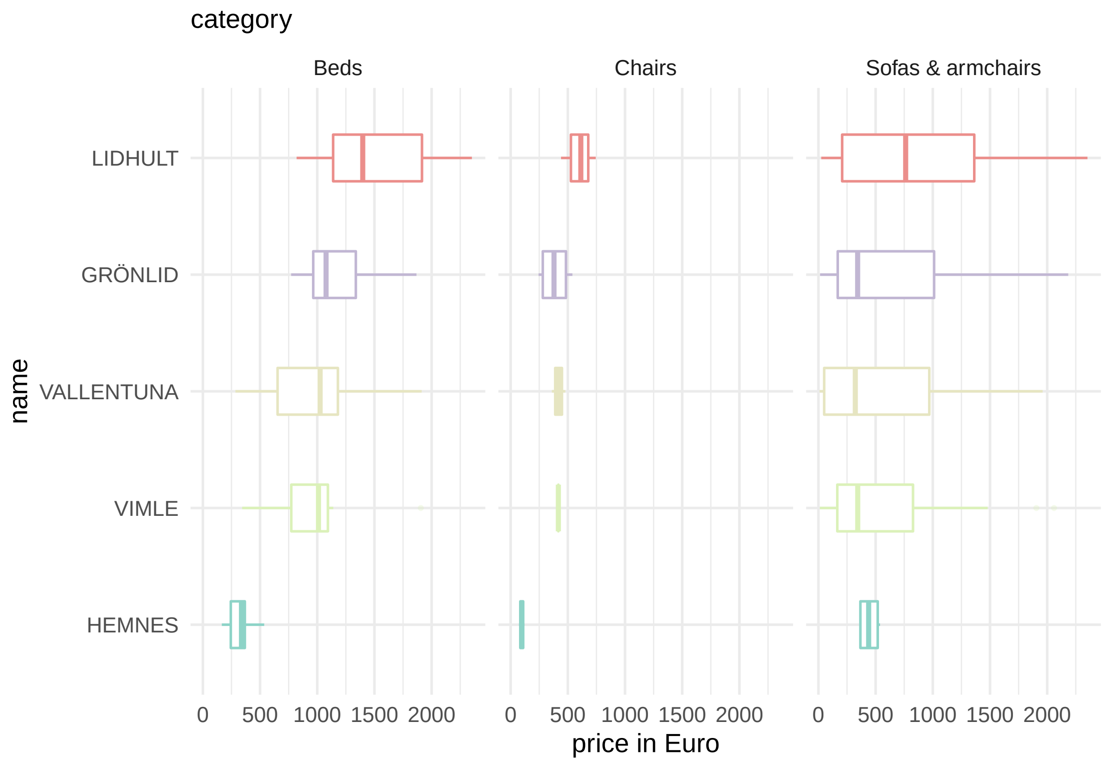

Figure 3.1: Homogeneity of category for selected combinations of name and category

3.2.3 Step 3: Missing Value Trouble

All numerical variables (`price`, `old_price` and `size_m3`) aren't arranged along a normal distribution (see 3.2), but rather follow an exponential decay (e^{-x}).

3.2.4 Step 4: Zeros

- many missing values from old price -> assume those were not on sale
- size_m3 missing values because of calculation formula
- designer -> removed values containing digits (were clearly falsely scraped)

3.2.5 Step 5: Collinearity between Independent Variables

- high collinearity between old price and price
- relatively high collinearity between price and size
- as can be seen in table

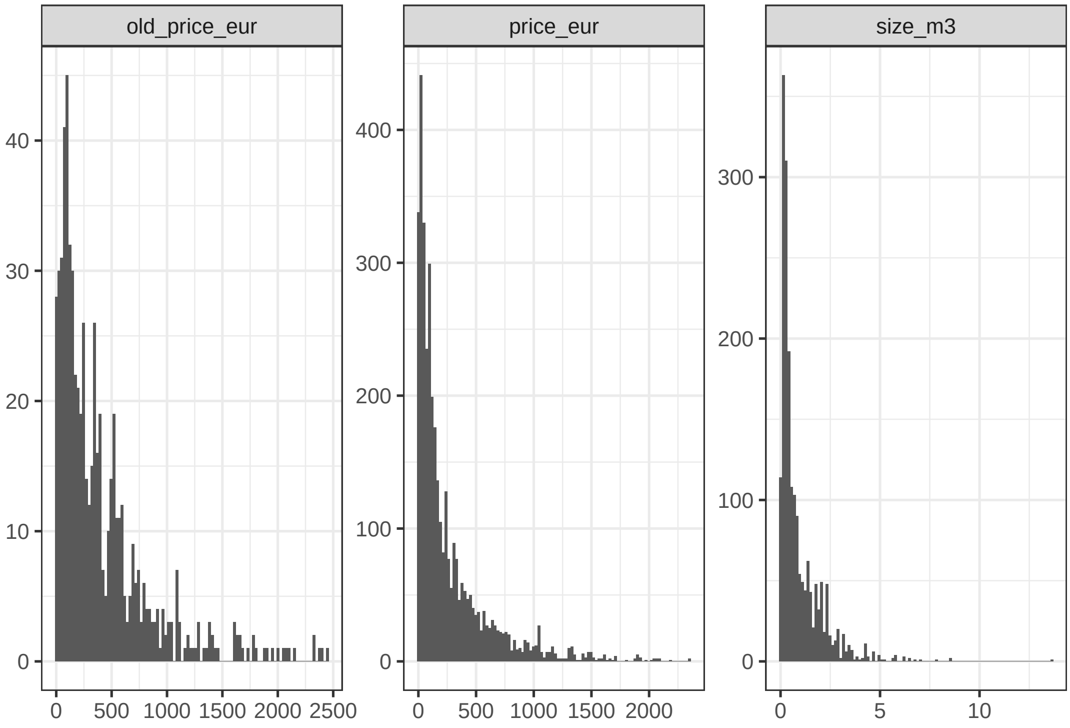

Figure 3.2: Homogeneity of category for selected combinations of name and category

3.2.6 Step 6: Relationship between Independent Variables and Price

- from eda_covariance.Rc (in Anhang + verweisen)
- strong relationship b/w price + old price & b/w price + size (see)
- other relationships aren't strong

3.2.7 Step 7: Interactions

- Coplotting designer and name works, while the two combination would not plot
- The linear model predicted infinite values and thus coplot the values properly for the other two options
- However, dropping all NA values and thus reducing the total data size to 354 observations would case for the combination coplot name and category while name + designer combination would not work

- The authors hypothesized that infinite values were caused by a division of zeros of the linear model since there occured 0 values in size
- This however proved to be wrong after applying respective filters
- The following interaction could be analyzed
- - There is probably no significant interaction between size, price, name & designer as can be seen in coplot -> lines are nearly parallel
- Based on the coplot of category and name with the 354 observations, inparallelity could be observed and thus could conclude a interaction. However, this could also be due to the small sample size
-

(footnote): the authors would highly appreciate any solutions on this matter

3.2.8 Step 8: Independence of Price

- Durch das tidying in (cross reference) duplicate removal -> Zuur paper Step 8 1. citation: meaning that information from any one observation should not provide information on another after the effects of other variables have been accounted for. This concept is best explained with examples.

3.3 Random Forest Regression Model

- We chose rf
- random forest erklären -> article (for in depth reference see article)
- why random forest (not lm)? -> article to explain why random forest is great to explain feature importance
  - see article
  - TODO: is normal distribution relevant for random forest (step 3) -> see article or site something

- To reproduce our results… (**Johannes**)
  - We chose randomForest R package for our analysis -> as in … (cite article)
  - data base is tidy ikea (step …)
  - then transform this data set to apply rf (johannes)
    - remove old price because of high correlation factor (step 5 + 6) which would fuck up our overall result
    - fct_lump erklären
    - rf has problems with na values -> 3 different methods to solve problem -> calculated 3 feature importances -> mean(a, b, c)

4
Results

- x-reference theoretical background: rf / variable importance
- describe plot comprehensively – w/ numbers from table

5 Discussion

5.1 size_m3

- material cost
- big items more expensive than small ones
- high correlation to price

5.2 designer

- designers with high number of products produce products in wide price range : plot reference -> appendix

- many different combinations of designer (49-53)

- many designer-combinations with low number of products:

    - n combinations with occurrences<5

-> overfitting: cite scholarly article -> low generalization : model might perfom bad on other data

5.3 name

- overfitting

5.4 category

- beds are more expensive than chairs : some categories are more expensive than others (show plot)
- but, price range varies heavily in category (show plot)

5.5 other_colors

- products of every price range have both options : low importance
- mean price is higher if other_colors is true (show plot covariance price+other_colors)
- interquartile-range is smaller if other_colors is false

5.6 sellable_online

sellable true: price range too high sellable ntrue: very rare

5.7 Conclusion & Ausblick

- Further research:
  - analyze designer overfitting, recommend using overfitting techniques (e.g. . . .)
  - analyze same research question with other techniques (lm, name more) and compare
  - data scraping from other country-pages

6 Individual Statements

6.1 Philip Krück

6.2 Johannes Pein

Appendices

A
Plots

A.1 Plot xyz

A.2 Plot abc

B
Another Appendix

Works Cited

Lain Zuur, E.N. Ieno (2010). "A protocol for data exploration to avoid common statistical problems". In: *British Ecological Society.* DOI: 10.1145/1738826.1738829.