Analyzing the Feature Importance of Different Variables on the Price of Ikea Products

Philip Krück, Johannes Pein

Hamburg School of Business Administration

B.Sc. Business Informatics (18A-BI)

Digital Toolbox: Data Business

Lecturer: Ulf Köther

Group Number: 7

Matriculation Numbers: 3938 (P.Krück), 4001 (J.Pein)

04.12.2020

Contents

Appendices

List of Abbreviations

R	Statistical Programming Language
MSE	Mean Squared Error
IQR	Interquartile Range

1 Introduction

This project report is an examination at the Hamburg School of Business Administration in the module ‘Data Business’ as a part of a Bachelor of Science degree program. The students were given a data set and the task was to first explore the data and then choose a research question which was to be answered scientifically with the help of the statistical programming language R. The authors of this report were given a data set which contains Ikea products and different features, such as price, dimensional measures, name, category and designers of the product.

2
Theoretical Background & Research Question

2.1 Data Set

by P. Krück 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 distinct observations after removing duplicates.

2.2 Theoretical Background

2.2.1 Random Forest Basics

by J. Pein

In order to analyze the feature importance in relation to the price variable, a random forest regression model was chosen. A random forest consists of many decision trees, which make a prediction based on a majority decision process. In standard decision trees, each nodes is split to achieve the best performing model. In random forests however, the nodes are randomly split. Compared to linear regression models, the random forests model not only takes into account the mean and covariance structure of response, but also deeper deeper aspects of data[1] resulting in a more advanced and robust model. To learn more about random forests, please see Breiman.[2]

2.2.2 Feature Importance

by J. Pein

There are different ways to measure feature importance. In this analysis permuting feature importance by a random forest algorithm is used. The algorithm leaves one feature out while leaving all others unchanged, calculating the mean squared error (MSE) of the predictions. This is done for every feature and tree, calculating the overall MSE of the whole model..[3]

2.3 Research Question

by P. Krück This paper explores the following research question:

How important are the different features of Ikea products in regard to their price?

The motivating forces for this research question are the possible implications for price determination of new items.

[1] Grömping, “Variable Importance Assessment in Regression: Linear Regression versus Random Forest.”

[2] “Random Forests.”

[3] Ibid.

3
Methods

3.1 Data Cleaning and Transformatoin

by P. Krück 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	FALS
2	9333523	NORDVIKEN / NORDVIKEN	Bar furniture	20950	No old price	FALS
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

by P. Krück The following sections explore our data based on the eight step data exploration protocol proposed by Zuur et al.[1]

3.2.1 Step 1: Outliers in Price and Independent Variables

by P. Krück 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

by P. Krück 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 (see fig. 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 ($2^3 = 8$).

[1] Zuur, "A protocol for data exploration to avoid common statistical problems."

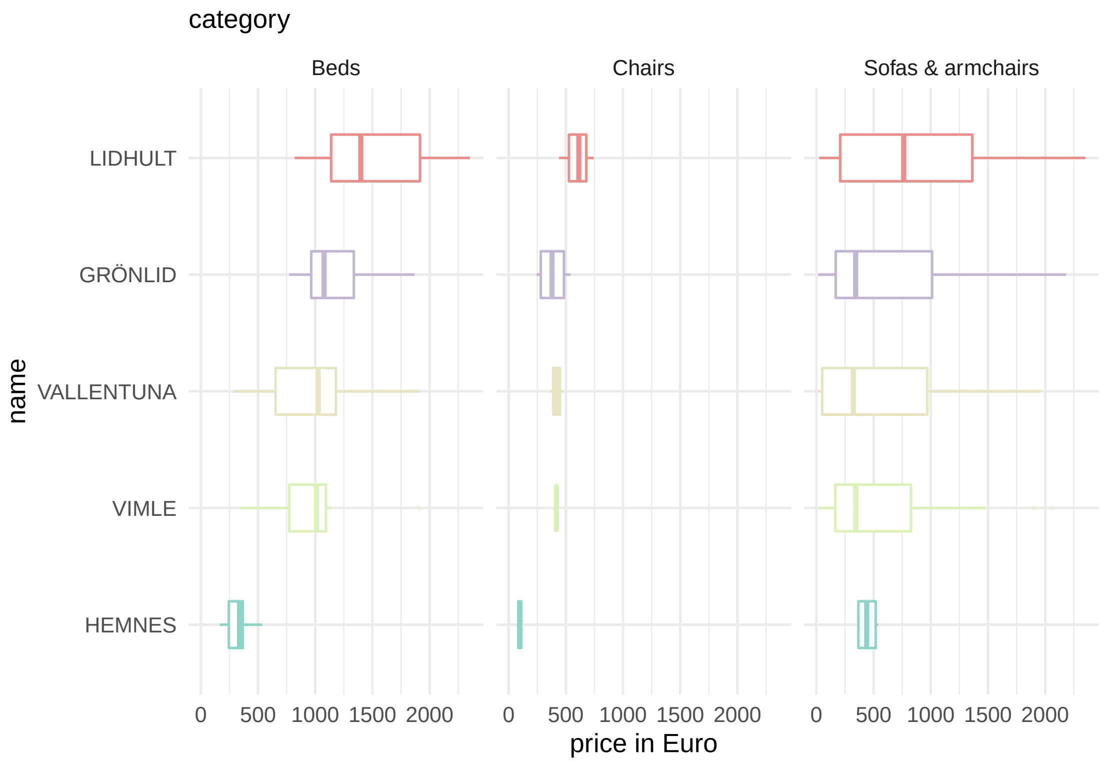

Figure 3.1: Homogeneity of category for selected combinations of name and category

3.2.3 Step 3: Normality

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

3.2.4 Step 4: Missing Values

by P. Krück All variables were examined for missing values. Only `designer`, `size_m3` and `old_price_eur` have missing values of 3.44%, 45.9% and 81% respectively (see fig. 3.3)). The missing values for designer were deliberately set to `NA` by the authors in the case where the values contained digits, which is clearly a scraping error. The `NA` values for the size can be explained due to the computation of this column vector. `size_m3` is the product of `depth`, `width` and `height`. If one of those three values is missing, the end result is also a missing value. The abscence of the old price variables is due to the fact that most items aren't on

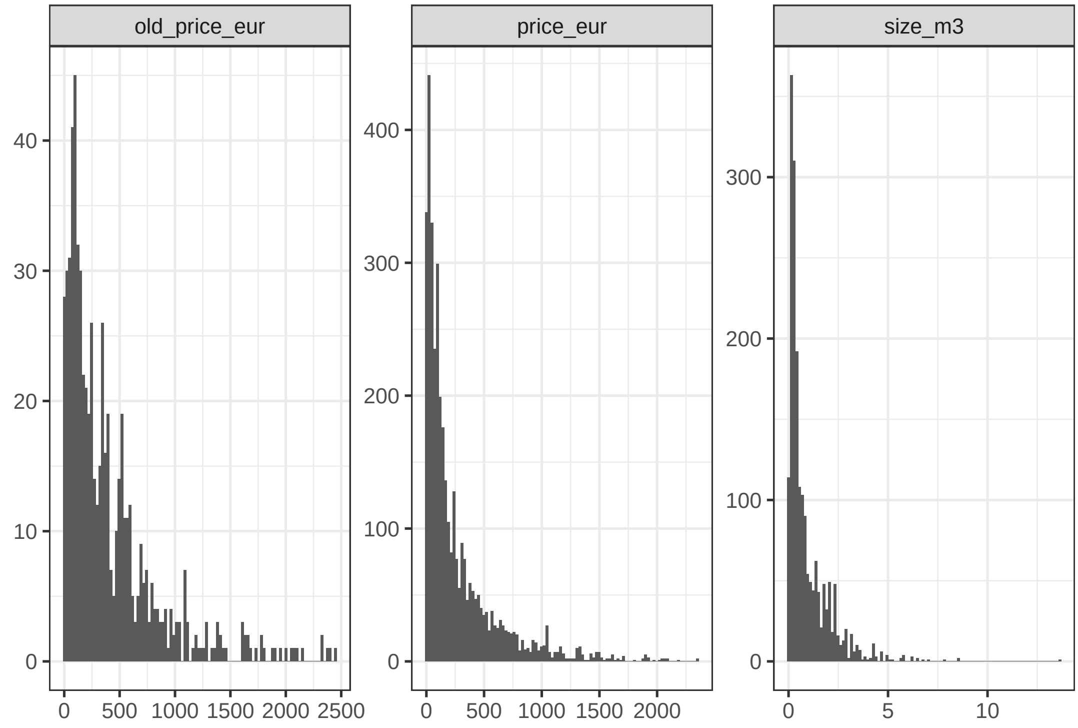

Figure 3.2: Homogeneity of category for selected combinations of name and category

Table 3.3: Initial Data Set formatting.

x
1
3
4

sale and thus don't have a missing value.

3.2.5 Step 5: Collinearity between Independent Variables

by P. Krück The old price has a rather high VIF which corresponds to high multicollinearity (see table . . .). Contrarily, size has a low VIF which translates to low multicollinearity among the other independent variables (see table . . .).

TODO: use values from step 5 of eda protocol

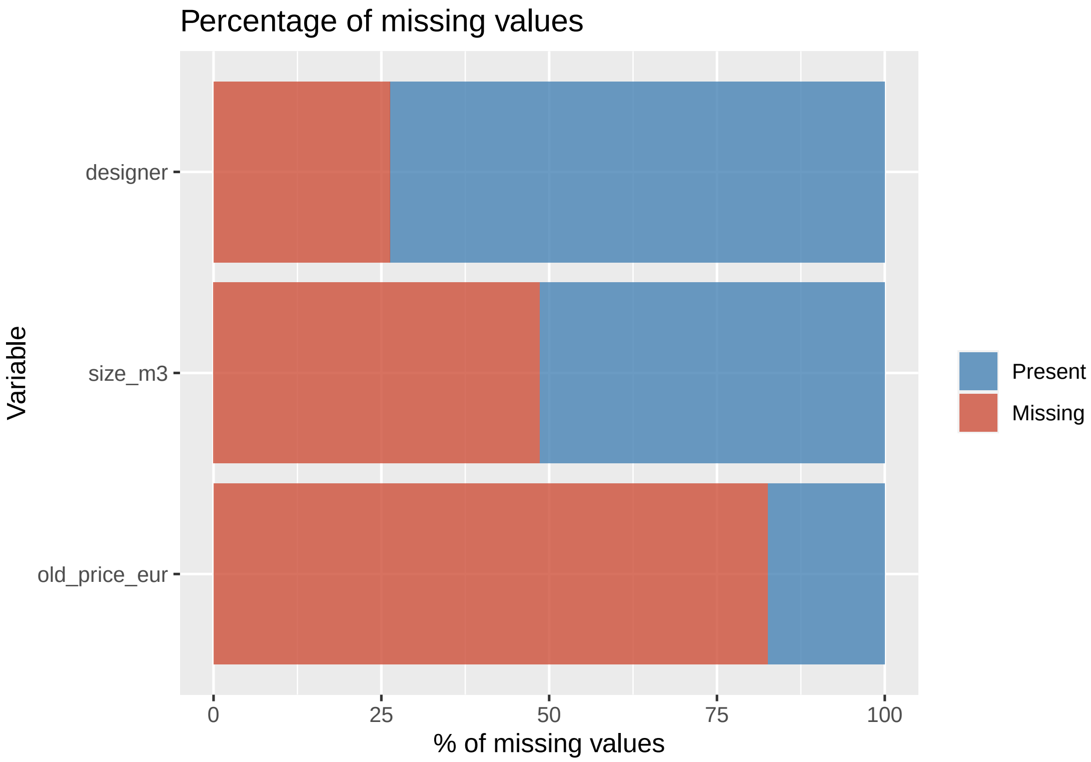

Figure 3.3: Homogeneity of category for selected combinations of name and category

3.2.6 Step 6: Relationship between Independent Variables and Price

by P. Krück Inspecting the relationship between the independent variables and price strong relationships between `old_price_eur` and `size_m3` can be observed, while no relationship can be observed for the other variables (see A.5). `old_price_eur` has a linear relationship (see fig. 3.4) whereas `size_m3` fits a second order polynomial (see fig. 3.5) to price.

3.2.7 Step 7: Interactions

by P. Krück The interactions between different variables is explored by the means of coplotting. Using this form plotting the relationship of two numerical variables is explored by creating a matrix of plots subdivided by two categorical variables. In the given data set there are three numerical and three categorical variables which can be explored in this form of interaction. For the numerical variables,

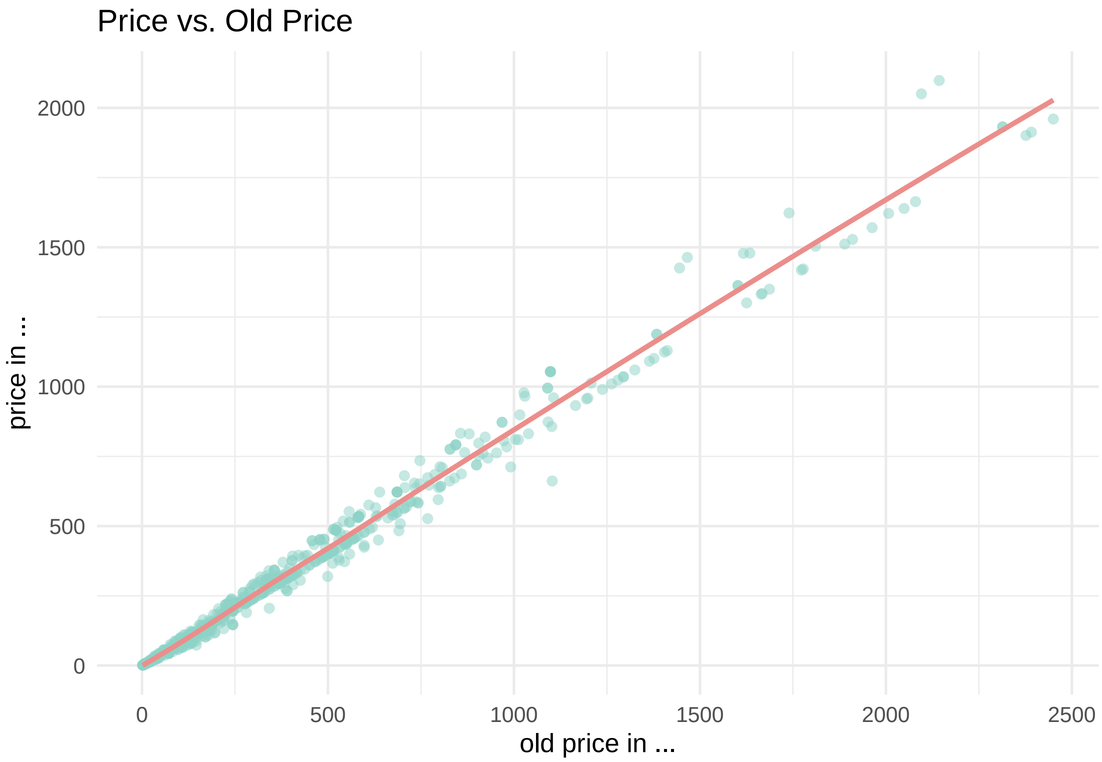

Figure 3.4: caption

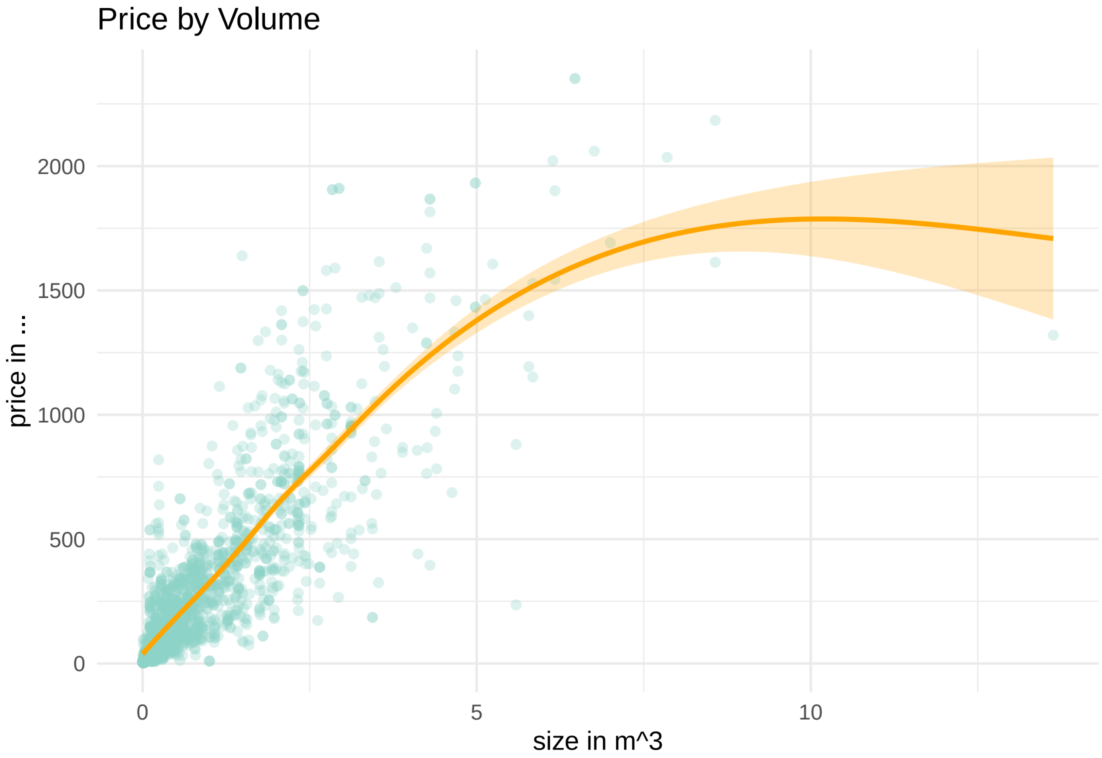

Figure 3.5: caption

`old_price_eur` has such a strong relationship with price (see . . .) that a more detailed breakdown by the categorical variables wouldn't reveal new information. This leaves the exploration of `size_m3` and `price_eur` broken down by designer, name and category resulting in (3 out of 2 = 3) combinations of coplots.

Interaction of size and price coplotted by designer and name

by P. Krück - 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

Problems with Coplot Development

by P. Krück This section describes a programming error the authors ran into regarding coplotting.

Unfortunately, the authors of this papers weren't able to fully explore all combinations. Plotting designer and name works (results described in above section) while the other two options would not plot properly. The authors could not fully debug the problem with these plots. The linear model predicted infinite values for some of the coplotted combinations for both combinations that wouldn't render correctly. Dropping all `NA` values left 354 observations and the coplot would correctly render for the combination of `name` and `category` while for `name` and `designer` it would not. The number of observations for `name` and `category` is rather low which lead the authors to discard it as an insignificant research finding. Still there seemed to be infinite values outputted by the linear method. The authors hypothesized that those values were caused by a division by 0 of the internal algorithm mechanics. This however proved to be wrong after applying the respective filters. The code for the 3 plots can be viewed in Appendix section xxxxx. The inclined reader is encouraged to play with the code. Sending any hints or even a solution to fix the code and fully render the plots would be highly appreciated by the authors.

[2]

3.2.8 Step 8: Independence Observations of Response Variable Price

by P. Krück The independence of observerations of the response variables assumes that "[...] information from any one observation should not provide information on another after the effects of other variables have been accounted for."[3] The data cleaning step (see data cleaning) left the observed data set in a tidy format ^["There are three interrelated rules which make a dataset tidy:

Each variable must have its own column. Each observation must have its own row. Each value must have its own cell." @ref([4], p.149, ll. 4-7)] which implies that the observations are independent of eachother.

3.3 Random Forest Regression Model

by J. Pein

This analysis was conducted using the R randomForest package, which is based on the original Breiman and Cutler's Fortran code for random forest regression. To learn more about how random forests work or the randomForest package, see Liaw and Wiener[5] or #chapter2 #TODO. To reproduce the analysis conducted in this paper, the prepatory steps are described here. The following steps are based on an already cleaned ikea data frame which is described in 3.1. This data frame is then transformed further to be used with the randomForest package.

First, the variable `old_price_eur` is removed from the data frame, due to a very high correlation and relationship to the price variable analyzed in **??** and 3.2.6 Then, the designers and names, which are not part of the 50 `designers` and 49 `names` with the highest number of occurences, are grouped in the `other` value. This is because the `randomForest` method does not allow categorical variables with more than 53

[2] Any questions, hints or solutions may kindly be sent to philip.krueck@myhsba.de.

[3] **???**

[4] **???**

[5] "Classification and Regression by randomForest."

predictors. The last step is dealing with the missing values in the data. As described in `??`, there are missing values in the `size_m3` and `designer` variables. To use the `randomForest` method of the randomForest package on the data, those missing values are dealt with using three different approaches. In the first approach the rows with missing values are deleted, reducing the total number of rows by aproximately 50%. In the second approach the missing values are dummy coded with a value of -1000. The third approach uses the `na.roughfix = na.omit` argument, which is the built in way of the randomForest package to deal with missing values. After preparing the data, the `randomForest` method of the randomForest package is applied to the data with number of trees set to 2000 and importance set to `TRUE`.

```
randomForest(price_eur ~ ., rf_ikea, ntree=2000, keep.forest=FALSE,
importance=TRUE)
```

Then the `importance` method of the randomForest package is used to save the feature importances, which are computed by permuting feature importance, which is described in #chapter2 #TODO. The three different approaches of dealing with the missing values in the data set lead to different results, so the authors chose to calculate the mean result of the three approaches. The result of this analysis is presented in the following chapter.

4
Results

by J. Pein

TODO: set n_trees to 2000 before handing in TODO: The results sectionshould comprise all necessary calculations, including checking of assumptions (if applicable), which are thendiscussedinconnection with the research questionin the following section.

In this chapter, the results of the analysis of the feature importance of different predictor variables (features) on the response variable price of Ikea products are presented.

As described in 3.3, the feature importance was calculated using permuting feature importance of the randomForest R package. In this analysis, feature importance is derived from the percentage increase of the mean squared error (MSE) of the overall random forest regression model with the response variable `price_eur`. When the percentage increase of the MSE is higher, the feature is more important, accordingly, when the percentage increase of the MSE is lower, the feature is less important.

Thus, as can be seen in figure 4.1, the most important feature is `size_m3` with an increase of the MSE of 182%. The second, third and fourth most important features are `designer` with an increase of 120%, `name` with an increase of 114%

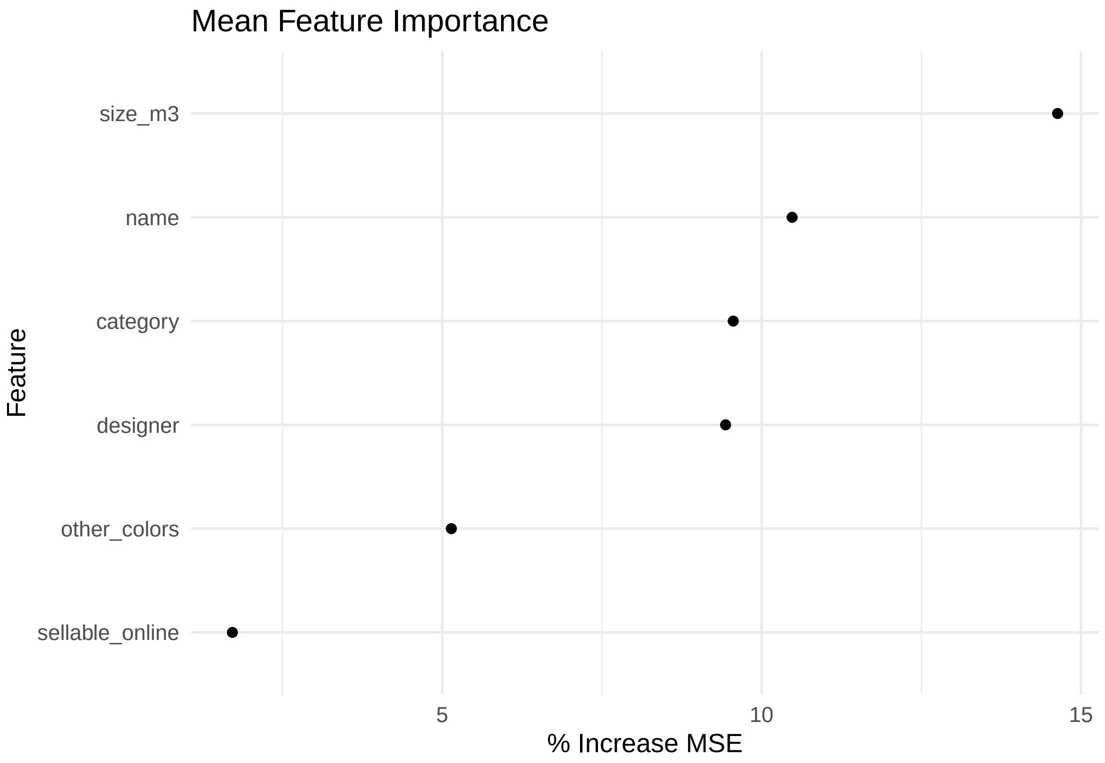

Figure 4.1: Mean Feature Importance

and `category` with an increase of 105%. The fifth most important feature is `other_colors` with an increase of the MSE of 78% and the least important feature is `sellable_online` with a 9% increase.

These results are further discussed in the following chapter.

5 Discussion

by J. Pein

In this chapter, the results are discussed in connection with the research question. The question that the results were supposed to answer is the following:

How important are the different features of Ikea products in regard to their price?

5.1 Feature Importance

The `size_m3` variable is the most important feature. Probably the main reason for this is that the size of a product is closely linked to its material cost. Big items are generally more costly to produce, thus leading to a higher selling price and vice versa. Due to the high correlation to the price variable described in **??** and 3.2.6, it is worth discussing where or not to include this variable in a possible predictive analysis model.

The `designer` variable is the second most important feature. It might seem, that this is due to *overfitting*, since the random forest regression model takes into account around 50 different combinations of designers, partly with a low number of occurences. But according to Breiman[1] random forest models are robust against overfitting. Further research should be conducted to analyze whether overfitting is

[1] "Random Forests."

present or not. When looking at the price distribution per designer, in can be clearly seen that the interquartile range (IQR) of price varies for each designer. In addition, the IQR often is smaller than 300€, thus showing a tendency towards a certain price range, which might be the reason for the relatively high feature importance of the designer variable. The plot @ref(price_distribution_designer) is available in the appendix. The random forest regression model includes around 50 different product names with partly small numbers of occurences. Thus, the relatively high feature importance of the `name` variable might also be caused by overfitting. As discussed above, further research should be conducted to analyze whether this is the case.

`Category` is another feature with a relatively high importance. This is because the different category's price distributions show a clear tendency towards certain price segments (see figure A.2), i.e. Wardrobes and beds are generally more expensive than chairs. To the authors it seems counterintuitive that `designer` and `name` have higher feature importances than the `category` feature, because the IQR in the price distribution per category is often a lot smaller than the IQR of the designer and name price distributions. This also hints towards overfitting of the name and designer variables. However, in the categories with the most occurrences, namely `Wardrobes`, `Sofas & armchairs` and `Beds`, the IQR is relatively large and there are many overlapping prices ranges for different categories, which explains a lower feature importance.

The feature importance of `other_colors` is the second lowest, but still considerable. This still relatively high feature importance might be due to the difference of the mean price and the relatively small IQR (see: figure A.3. On the other hand, there is a large overlapping area within the IQR in the two expressions of `other_colors` possibly reducing the feature importance.

The very low feature importance of the `sellable_online` variable is probably because the low number of occurences of a product being sellable online. Only around 0.6% of the products are sellable online.

5.2 Conclusion

Generally, the proposed random forest model was not validated with a test or validation data set. In further research, the results presented on the feature importance of different variables on the price variable should be validated by other techniques than used in this analysis to get unbiased results.

Also, the data was scraped from the arabian Ikea website online , thus this analysis mainly focusses on Ikea products in the Saudi Arabian market. To analyze the geographically independent feature importances, more data should be scraped from more international Ikea websites. The research question of this paper, *How important are the different features of Ikea products in regard to their price?*, could thus not be answered for the global Ikea product market, but for the Saudi Arabian market only.

Furthermore, based on this analysis a predictive model could be developed which predicts the price of Ikea products in the Saudi Arabian Market based on the feature analyzed. This could be used by Ikea internally to analyze if the price of their new product aligns with the prices of the currently available product.

6 Individual Statements

6.1 Philip Krück

6.2 Johannes Pein

Appendices

A
Appendix

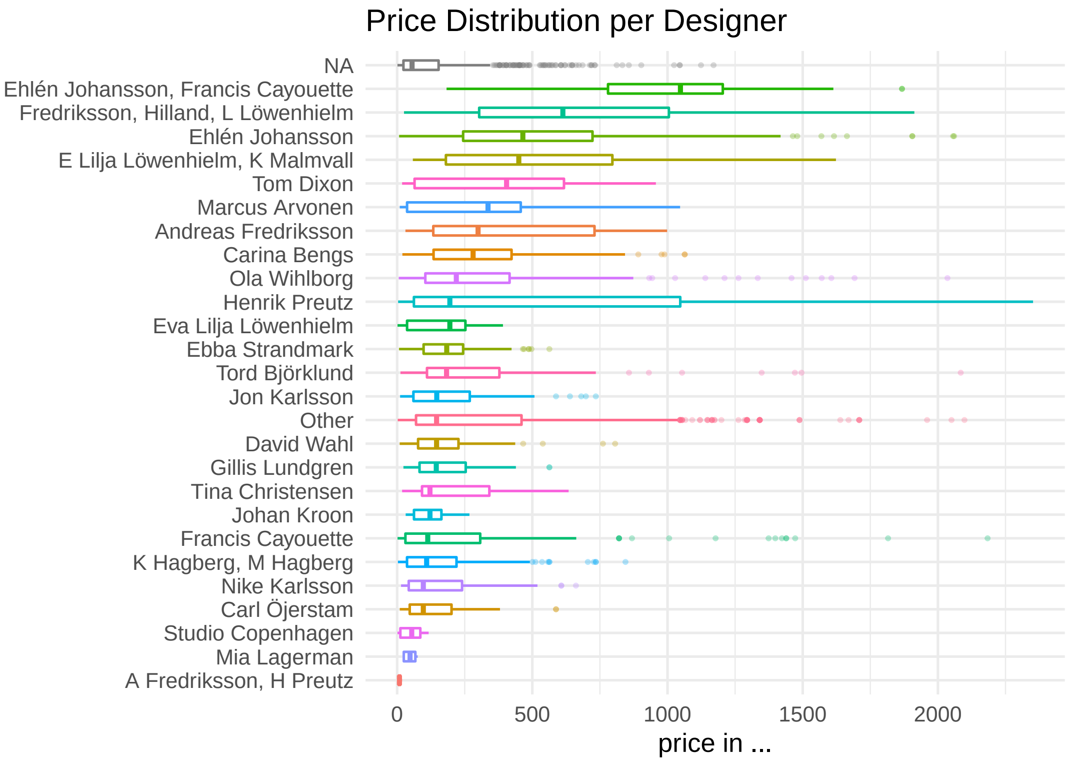

Figure A.1: Price Distribution per Designer

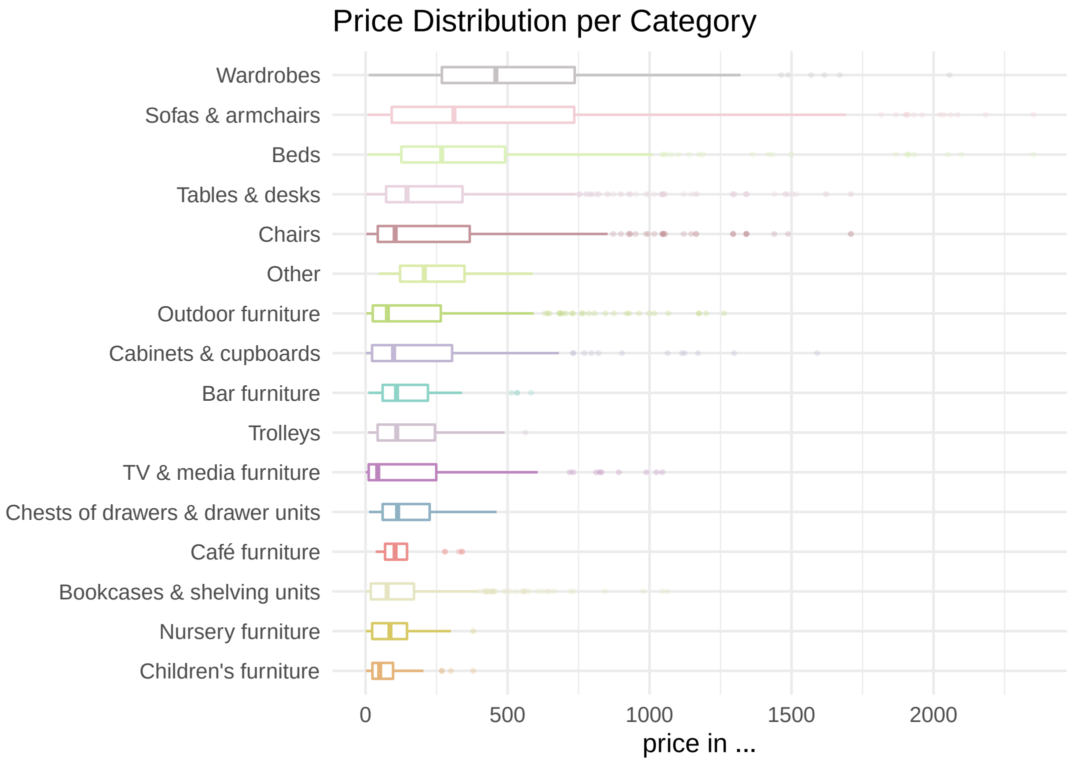

Figure A.2: Price Distribution per Category

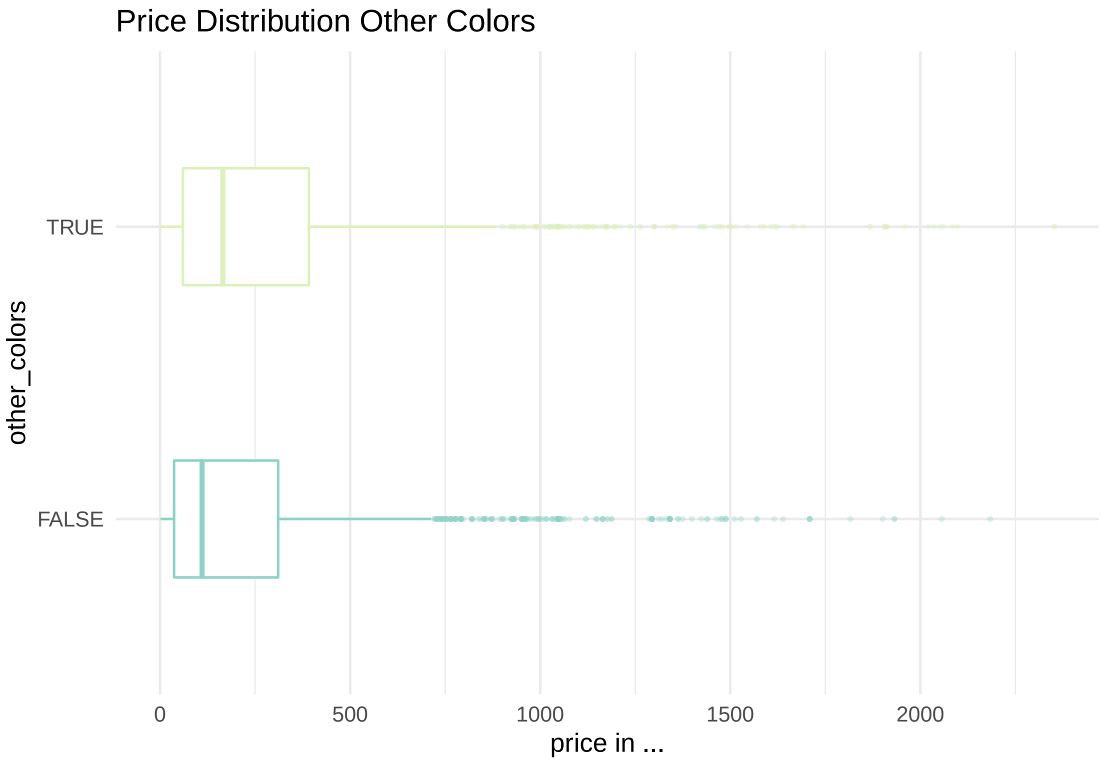

Figure A.3: Price Distribution Other Colors

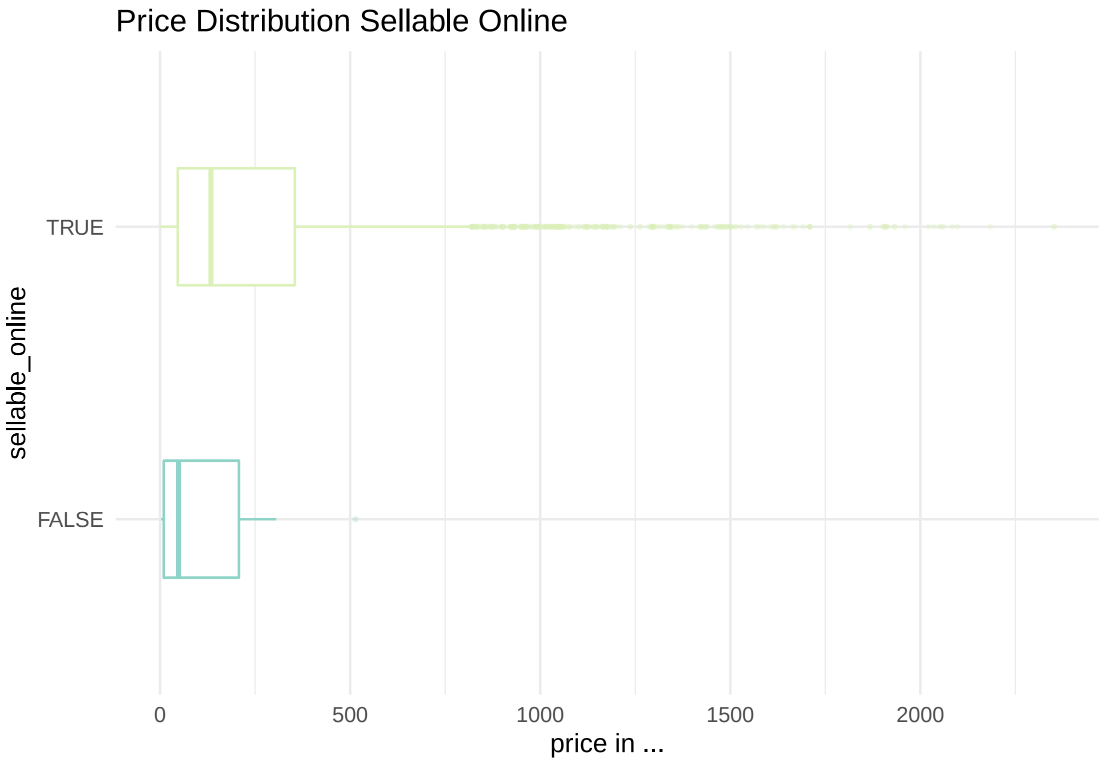

Figure A.4: Price Distribution Sellable Online

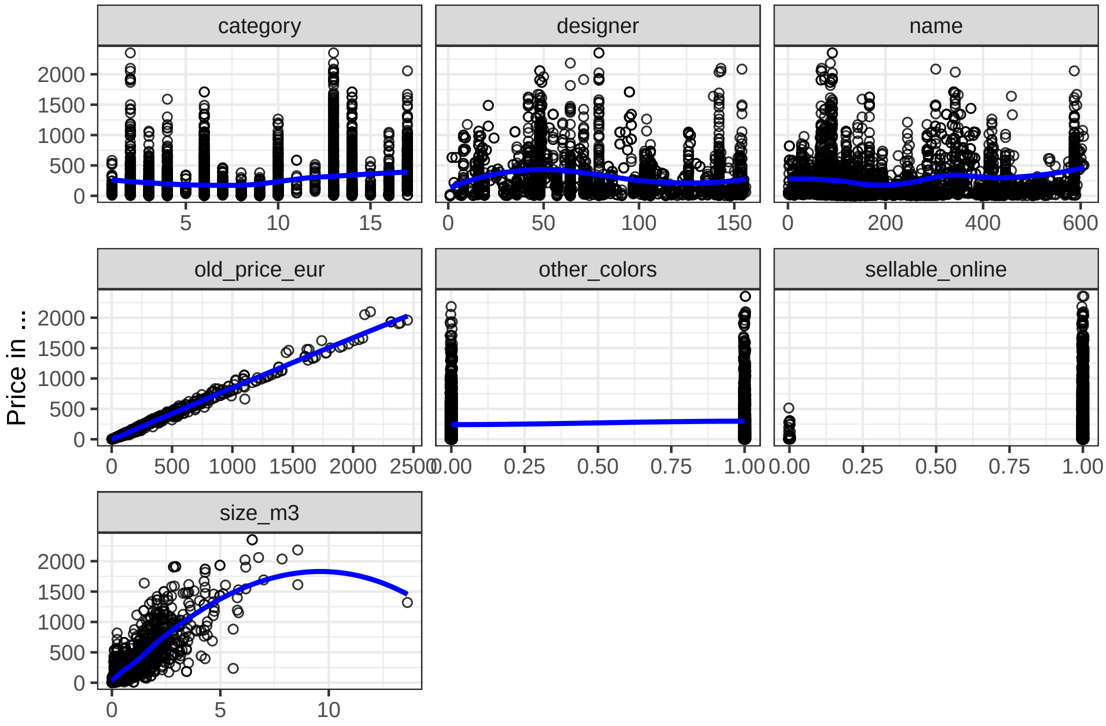

Figure A.5: Relationship between Independent Variables and Price

B
Bibliography

Breiman, Leo. “Random Forests.” *Machine Learning*, no. 45 (2001). https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf.

Grömping, Ulrike. “Variable Importance Assessment in Regression: Linear Regression versus Random Forest.” *The American Statistician*, nos. Vol. 63, No. 4 (2009): 308–19. https://prof.beuth-hochschule.de/fileadmin/prof/groemp/downloads/tast_2E2009_2E08199.pdf.

Liaw, Andy, and Matthew Wiener. “Classification and Regression by randomForest.” *R News*, nos. Vol. 2/3, December 2002 (2001). https://www.researchgate.net/publication/228451484_Classification_and_Regression_by_RandomForest.

Zuur, E.N. Ieno lain. “A protocol for data exploration to avoid common statistical problems.” *British Ecological Society*, 2010. doi:10.1145/1738826.1738829.