HAMBURG SCHOOL OF
BUSINESS ADMINISTRATION

Digital Tool Box Data Business – Project Report

Project work as part of the Bachelor of Science (B.Sc.) in Business Informatics

Benedikt Kronhardt (5089)
Börge Meyer (5076)

Project Theme	Cost of Living Index: Does the classification of a country as a developing or industrialized country have a significant impact on the cost of living index?
Study Group	20A-BI2
Lecturer	Ulf Köther
Submitted	December 6, 2022
Team Number	1
Wordcount	2754

Contents

List of Figures

1 Introduction

The standard of living became more and more important for the world population. But every standard of living comes at a price. How high the standard of living is in a country can be analyzed and compared between countries with the help of the cost of living index.

1.1 Task description

Our task was to analyze a data set and write a report about it using R, RStudio, RMarkdown and the procedures of literate programming to put together a PDF-manuscript. In our team we have received the "cost of living" data set and analyzed it with the research question "Does the classification of a country as a developing or industrialized country have a significant impact on the cost of living index?"

1.2 Structure

Our document is divided into five chapters. This chapter is the introduction, where we describe the task and how we imported our dataset, respectively added other datasets. In the chapter "Theoretical Background" the theoretical background of the dataset and the data collection is briefly discussed, before in the methods section the dataset is statistically described, including information on the variables' distribution, missing values, categories and the relationships between the variables. The results section comprise all necessary calculations, which are then discussed in connection with the research question in the following section ("Discussion").

1.3 Setup

After the required libraries, which will be worked with in the following, were installed, the libraries still had to be imported in order to be able to use them.

```r
library(tidyverse)
library(dplyr)
library(stringr)
library(ggplot2)
library(maps)
library(janitor)
library(modelsummary)
library(car)
library(carData)
library(gpairs)
library(GGally)
```

Subsequently, the data had to be read in. This could be initialized with the following command, after the data set was added as a csv file in the folder “02-data”. To be able to work better with the names of the columns and the dataset in general, the command “janitor::clean_names” was executed. With this, for example, the spaces were removed and the names were all written in small letters.

```r
costOfLiving <- read_delim("02-data/cost-of-living-2017.csv",
                           delim = "\t", escape_double = FALSE,
                           trim_ws = TRUE)
costOfLiving <- janitor::clean_names(costOfLiving)
```

To make it easier to split the data by region, we imported a csv file that shows the names of the countries in this world and their corresponding regions.

First we had to import the dataset, which we named “continents”. This dataset is from the website “kaggle”, named “Country Mapping - ISO, Continent, Region”. [1]

```r
#import list of continents and countries
continents <- read_csv("02-data/continents2.csv")
continents <- janitor::clean_names(continents)
```

[1] Kaggle (2019)

To be able to do a join with the raw data, we had to rename the column "name" to "country". After that, a left join could be performed on the renamed column. Since we only needed the column "region", a select for this one column was performed within the join.

```
#rename attribute 'name' in 'country' to perform a left join.
continents <- rename(continents, country = name)
costOfLivingAndContinents <- left_join(costOfLiving, select(continents, country, region),
```

Now it was possible to check if a country was not assigned to a region. Since the country Kosovo could not be assigned to a region, this had to be done manually.

```
costOfLivingAndContinents[486, 12] <- "Europe"
```

To assign the different countries in our dataset to either a developing or an industrialized country, we also imported a new csv file, which we named "dd". We created this file ourselves, based on data from UN (2014).

```
#import list of Countries and development status
dd <- read_delim("02-data/developed_and_developing_countries.csv",
                 delim = ";", escape_double = FALSE,
                 trim_ws = TRUE)
dd <- janitor::clean_names(dd)
```

To format the category as a double value, we executed the following commands.

```
dd$category[dd$category == "developed"] <- 1.0
dd$category[dd$category == "developing"] <- 0.0
dd$category <- as.double(dd$category)
```

Once this was done, we scanned the dataset for various capitalization errors and corrected them. Also we have renamed the column category to development.

```
dd$country[dd$country == "italy"] <- "Italy"
dd$country[dd$country == "Hong Kong SAR"] <- "Hong Kong"
dd$country[dd$country == "Taiwan Province of China"] <- "Taiwan"
dd$country[dd$country == "Russian Federation"] <- "Russia"
dd$country[dd$country == "Viet Nam"] <- "Vietnam"
dd$country[dd$country == "Bosnia and Herzegovina"] <- "Bosnia And Herzegovina"
dd$country[dd$country == "Kosovo"] <- "Kosovo (Disputed Territory)"
colnames(dd)[2] <- "development"
```

In the end, we were able to perform a left join and thus add the categorization of development countries to our dataset.

```
dataFinished <- left_join(costOfLivingAndContinents,dd, by="country")
```

2 Theoretical background

2.1 Cost of Living Index

Because of different prices, living standards, currencies and other factors, it is not possible to compare the cost of living in different countries properly.

To be able to compare the cost of living between different countries, the Cost of Living Index is used - also abbreviated as CLI in the following. The cost of living is the financial resources needed to cover, in a given place and in a given period of time, the basic expenses for a given standard of living, such as a shelter, food, medicines and others. The CLI enables the comparison of expenditures between different places in the world and at different times in history.[1]

In economics, the cost-of-living index describes the ratio of the minimum expenditure required to achieve a given indifference curve between two prices. The calculation not only requires two different price groups, but is also dependent on a preference order of the required living goods and on a basic indifference curve describing the utility of two products. Among the two prices needed, e.g., from two different places, one is called the comparison price and the other is called the reference price or the base price. The base price is then used to illustrate on which prices the Cost-Of-Living Index is based and calculated. The calculated index is then dependent on the comparison prices determined. Further, the general logic of the cost-of-living index is best understood when the index is interpreted in the multiple context of temporal and spatial comparisons.[2]

2.2 Industrialized, emerging and developing countries

In general, countries are divided into industrialized, emerging and developing countries. States in which the economic performance is supported by a large part of the resident companies are referred

[1] Banton (2021)

[2] Pollak (1989)

to as industrialized countries. Such countries stand out due to their high per capita income, which results from the available standard of education, high productivity in production, good external trade relations and usually a currency with low inflation.[3]

A country that is in the process of becoming an industrialized country is called an emerging country. These are nevertheless referred to the category of developing countries. Emerging countries are identifiable by their above-average economic growth. Nevertheless, emerging countries are similar to developing countries in the social structure, such as in the level of education, mortality and access to infrastructure.[4]

The third category is developing countries, which are associated with poor food supply, high poverty, poor health care and educational opportunities. In association with the characteristics, such countries have an overall low standard of living and a preponderance of labor in agriculture and external economic difficulties.[5]

To analyze the available data, developing countries were combined with emerging economies and contrasted with developed countries.

[3] BPB (2021)

[4] BMZ (n.d.b)

[5] BMZ (n.d.a)

3 Methods

3.1 Data Description

The provided data consists of 511 different datasets from 110 different states. The data was set up into City, State, Country, Cost of Living Plus Rent Index, CLI, Rent Index, Groceries Index, Restaurant Index, Local Purchasing Power Index, Leverage Model 1 and Leverage Model 2 attributes.

In Figure 3.1 can be seen how many datasets are available per region.

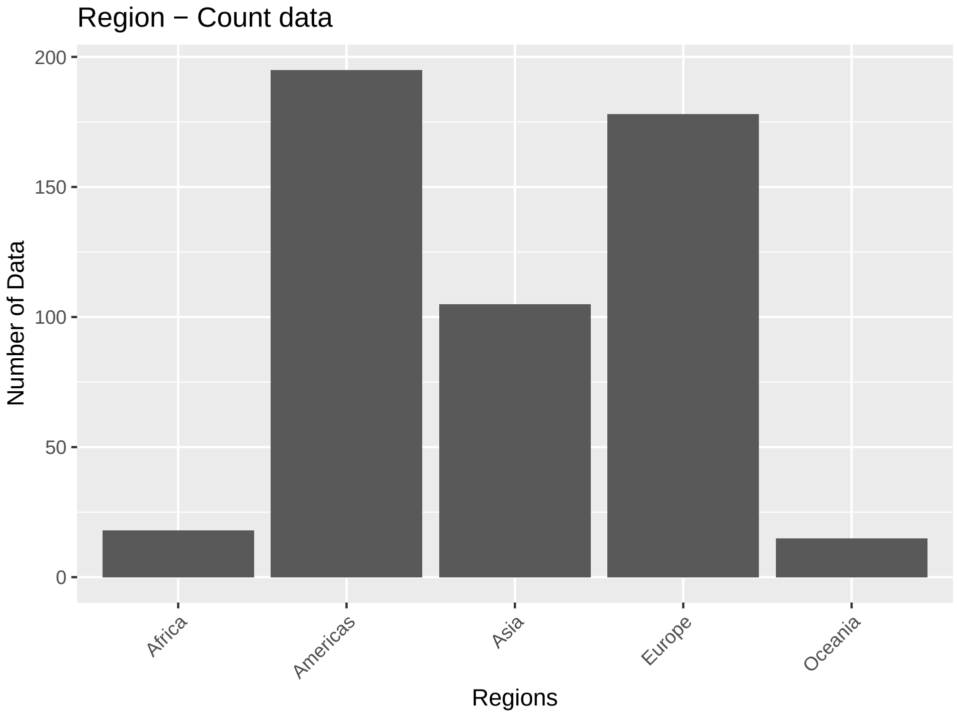
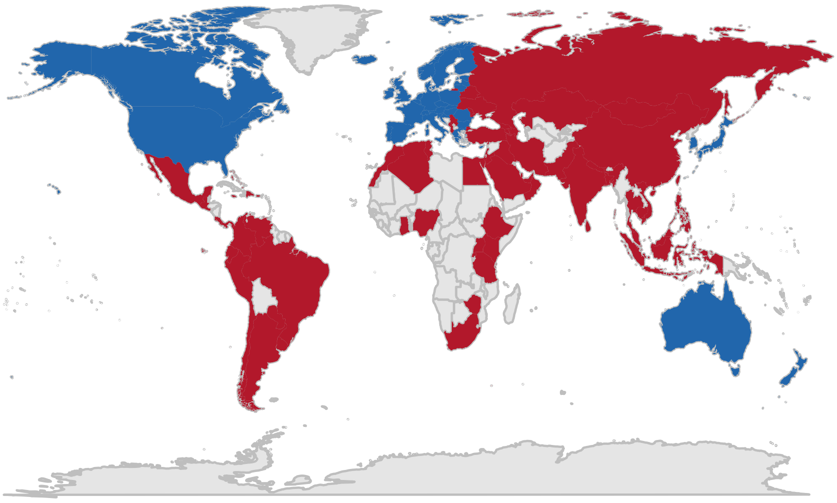


Figure 3.1: Count of Data from different regions

In addition to the overview of datasets by region in figure 3.1, a world map in figure 3.2 has been created to illustrate the countries from which the datasets originate. The data sets from the industrialized countries were marked in blue and those from the emerging and developing countries in red.



Red = Developing Countries, Blue = Industrialized Countries

Figure 3.2: Industrialized and developing countries

3.2 Exploratory Data Analysis

First of all, we had to check, if there are missing values inside of the data-set. Therefore we used the following code to proof this:

```
summary(is.na(dataFinished))
```

```
#>     city           state          country
#>  Mode :logical   Mode :logical   Mode :logical
#>  FALSE:511       FALSE:128       FALSE:511
#>                  TRUE :383
#>  cost_of_living_plus_rent_index    cli
#>  Mode :logical                  Mode :logical
#>  FALSE:511                      FALSE:511
#>
#>  rent_index      groceries_index restaurant_price_index
#>  Mode :logical   Mode :logical   Mode :logical
#>  FALSE:511       FALSE:511       FALSE:511
```

```
#>
#>  local_purchasing_power_index leverage_model_1
#>  Mode :logical                Mode :logical
#>  FALSE:511                    FALSE:511
#>
#>  leverage_model_2   region        development
#>  Mode :logical    Mode :logical   Mode :logical
#>  FALSE:511        FALSE:511       FALSE:511
#>
```

As it can be seen, there were 383 missing values inside the column “state”. However, since the column has no bearing on our research question, we decided to disregard this column. With the city column we have a more meaningful basis to answer our question. To disregard this column, we cut it off. To do this, we used to following R code chunk. Because it is the second column, we can just delete this column.

```
dataFinished <- dataFinished[-2]
```

We also truncated the leverage_model_1 and leverage_model_2 columns, since we did not work with these columns any further.

```
dataFinished <- dataFinished[-9]
dataFinished <- dataFinished[-9]
```

To determine if outliers exist within the data set, we chose to draw a boxplot.

As can be seen from the figure 3.3, there are several outliers within the data set. In order not to distort the result, we decided to keep these outliers and to continue working with them.

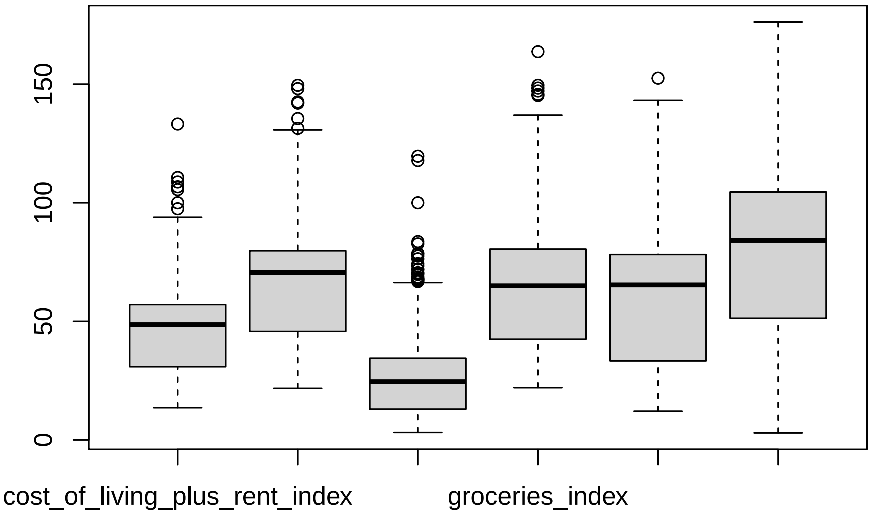


Figure 3.3: Boxplot of Data to identify outliers

4 Results

To determine whether there is a significant difference between developing and developed countries, we decided to run a multiple linear regression. This is to determine whether the classification into a developing country has a significant influence on the cost of living index or not.

4.1 Multiple linear regression

Within multiple linear regression, our dependent variable (y) is the cost of living index. Our independent variables (x) are the rent index, the groceries index, the restaurant price index, the local purchasing power index and the development status.

In order to perform a multiple linear regression, some conditions have to be fulfilled, which we will check in the following.

First, there must be a linear relationship between the x variables and the y variable. Also, the y variable must be metrically scaled, which is given.

Third, the residuals should be approximately normally distributed. We proved this graphically with the help of a histogram.

First, we need to set up our model.

```
model <- lm(cli ~ rent_index + groceries_index + restaurant_price_index
            + local_purchasing_power_index + development, data = dataFinished)
```

After that, we can create a histogram from our model.

From the histogram we can see that the distribution can be considered normally distributed, therefore this condition is also fulfilled. Scaling is also given, since the cost of living index is on a scale.

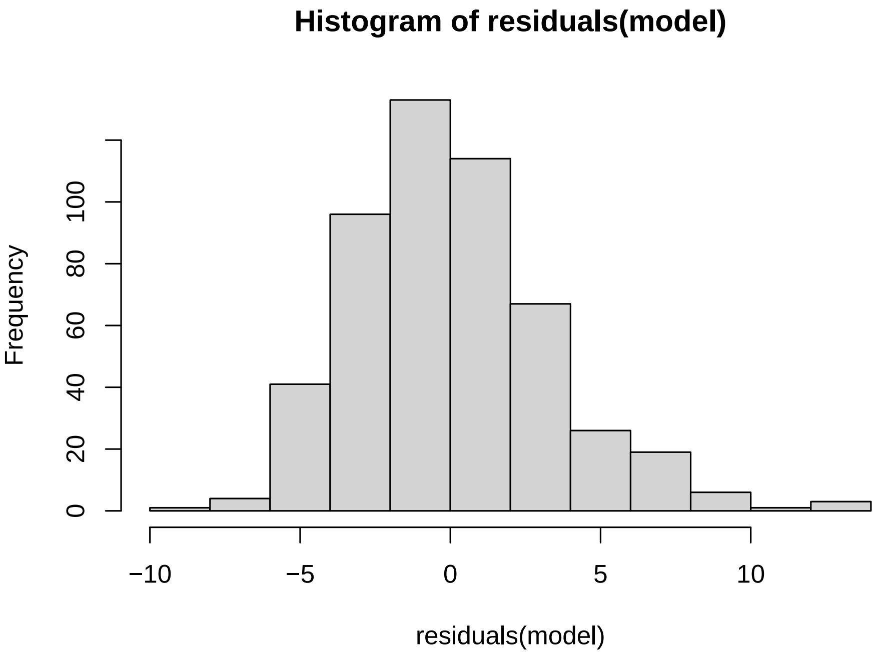


Figure 4.1: Histogram of the multiple linear regression model

The last condition we checked is that there must be no multicollinearity within the independent variables. To check this, we created a correlation matrix. First, we generated a subset from the data in which the variables to be tested are stored. Then we created the correlation matrix from this subset and worked with the pearson method.

```
#>                               rent_index groceries_index
#> rent_index                     1.0000000       0.7674361
#> groceries_index                0.7674361       1.0000000
#> restaurant_price_index         0.7523090       0.8518550
#> local_purchasing_power_index   0.6000432       0.6458339
#> development                    0.4730879       0.5998356
#>                               restaurant_price_index
#> rent_index                                 0.7523090
#> groceries_index                            0.8518550
#> restaurant_price_index                     1.0000000
#> local_purchasing_power_index               0.6436926
#> development                                0.6838520
#>                               local_purchasing_power_index
#> rent_index                                       0.6000432
#> groceries_index                                  0.6458339
#> restaurant_price_index                           0.6436926
#> local_purchasing_power_index                     1.0000000
#> development                                      0.6425433
#>                               development
#> rent_index                      0.4730879
#> groceries_index                 0.5998356
#> restaurant_price_index          0.6838520
#> local_purchasing_power_index    0.6425433
#> development                     1.0000000
```

Since the correlation between restaurant price index and groceries index is $0.851855 >$ 0.8, this may indicate that there is multicollinearity. To confirm this, we used another method to check for multicollinearity, the method of Variance Inflation Factor values.

```
#>                   rent_index              groceries_index
#>                     2.821033                     4.346802
#>       restaurant_price_index local_purchasing_power_index
#>                     4.863177                     2.187319
#>                  development
#>                     2.241055
```

Since according to this method none of the values is $>$10 we have rejected the theory of multicollinearity.

Now that all the assumptions can be accepted, we come to the actual evaluation of the model.

```
#> 
#> Call:
#> lm(formula = cli ~ rent_index + groceries_index + restaurant_price_index + 
#>     local_purchasing_power_index + development, data = dataFinished)
#> 
#> Residuals:
#>     Min      1Q  Median      3Q     Max 
#> -8.4054 -2.4098 -0.2694  1.8302 12.6586 
#> 
#> Coefficients:
#>                               Estimate Std. Error t value
#> (Intercept)                  10.785427   0.468691  23.012
#> rent_index                    0.028668   0.013976   2.051
#> groceries_index               0.479207   0.012561  38.152
#> restaurant_price_index        0.427674   0.012145  35.213
#> local_purchasing_power_index -0.028731   0.006438  -4.463
#> development                   0.466335   0.461661   1.010
#>                              Pr(>|t|)    
#> (Intercept)                   < 2e-16 ***
#> rent_index                     0.0408 *  
#> groceries_index               < 2e-16 ***
#> restaurant_price_index        < 2e-16 ***
#> local_purchasing_power_index 9.98e-06 ***
#> development                     0.3129
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 3.317 on 505 degrees of freedom
#> Multiple R-squared:  0.9782, Adjusted R-squared:  0.978
#> F-statistic:  4539 on 5 and 505 DF,  p-value: < 2.2e-16
```

The model makes a significant explanatory contribution, as the p-value is well below 0.05, and we can proceed with the interpretation of the further results.

As we can see, according to the p-values, all variables except the classification of development have a significant impact on the cost of living index.

5 Discussion

5.1 Critical assessment of the data

The objective of this study was to determine to what degree the status as an industrialized or developing country has an influence on the Cost of Living Index.

The critical review allows first of all to scrutinize the available data. Most of the data sets that were used as a basis for this work did not include all existing countries. In addition, it must be mentioned that a large number of African countries in particular are not included in the initial data. This could have biased the results of the work (see figure 3.2).

Furthermore, data were added that resulted in additional analysis possibilities, such as the representation of industrialized and developing countries. Data from the United Nations is considered to be trusted because the United Nations is an official and recognized organization.

Data produced by third parties are classified as less trustworthy, as this can lead to falsification. Since this was based on the regional allocation of the data provided for the different countries, the usage does not have a high weighting in the result.

5.2 Expressiveness of the model used

Furthermore, it must be critically questioned whether the multiple linear regression model used was really suitable to be applied to the research question. A multiple linear regression model was used with all numerical variables within the data set to determine which variables have a significant influence on the cli. Alternatively, the model could have been set up with only the variable developed and one other variable such as the rent index. However, we decided to use all variables in order to obtain the highest possible R^2. A simple linear regression between the cli and the developed variable was not possible because the linear relationship does not exist. In order to answer the research question more comprehensively, an additional t-test could have been implemented. However, we decided against this, as it would have exceeded the scope of our work.

5.3 Answer to the research question

Within the report, we conclude that the classification of a country as a developing or industrialized country does not have a significant impact on cli (see chapter 4). This is true for our data set. However, some countries are missing from our data set, so we cannot make this statement universally (see chapter 5.1).

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2022. *Rmarkdown: Dynamic Documents for r*. https://CRAN.R-project.org/package=rmarkdown.

Banton, Caroline. 2021. "Cost of Living: Definition, How to Calculate, Index, and Example." March 2021. https://www.investopedia.com/terms/c/cost-of-living.asp.

BMZ. n.d.a. "Entwicklungsland." Bundesamt für wirtschaftliche Zusammenarbeit und Entwicklung. https://www.bmz.de/de/service/lexikon/entwicklungsland-14308.

———. n.d.b. "Schwellenland." Bundesamt für wirtschaftliche Zusammenarbeit und Entwicklung. https://www.bmz.de/de/service/lexikon/schwellenland-14810.

BPB. 2021. "Industrieländer." Bundeszentrale für politische Bildung. June 2021. https://www.bpb.de/kurz-knapp/lexika/lexikon-der-wirtschaft/19720/industrielaender/.

Kaggle. 2019. "Country Mapping - ISO, Continent, Region." December 2019. https://www.kaggle.com/datasets/andradaolteanu/country-mapping-iso-continent-region.

Pollak, Robert A. 1989. *The Theory of the Cost-of-Living Index*. Oxford University Press.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

UN. 2014. *World Economic Situation and Prospects*. United Nations.

Declaration of Honor

We hereby declare that

1. we wrote this project report without the assistance of others;
2. we have marked direct quotes used from the literature and the use of ideas of other authors at the corresponding locations in the thesis;
3. we have not presented this thesis for any other exam. We acknowledge that a false declaration will have legal consequences.

Hamburg, December 6, 2022

Börge Meyer, Benedikt Kronhardt

We accept that the HSBA may check the originality of our work using a range of manual and computer based techniques, including transferring and storing our submission in a database for the purpose of data-matching to help detect plagiarism.

Hamburg, December 6, 2022

Börge Meyer, Benedikt Kronhardt