Roger Bär
Funkerstrasse 13
3013 Bern
roger.baer@cde.unibe.ch

Data Science Project

Sustainable Quality of Life: Conceptual Design Report

January 2020

ABSTRACT

In this data science project, the aim is to provide a first exploratory data analysis and to identify factors impacting life satisfaction of people. More specifically the objectives are: (1) to preprocess the collected data, (2) to describe the data, and (3) to identify correlations and factors impacting life satisfaction. For this purpose, a written survey that has been conducted in three nature parks (UNESCO Biospäre Entlebuch, Naturpark Gantrisch, Jurapark Aargau) and one control region will be used. Results show that life satisfaction is different according to different groups. However, the current regression results do not (yet) provide good results and needs substantial improvement. Further interesting analysis include the inclusion of additional variables and the statistical inference to the entire population.

Deliverables: Conceptual Design Report (this document) and GitHub Repository (https://github.com/bushroot/2020_ADS_M1)

TABLE OF CONTENTS

OBJECTIVE

Most European countries have a relatively high average quality of life. However, the high quality of life is often accompanied by a high consumption of resources, which contributes to global environmental and social problems. To this end, a concept for a "sustainable quality of life" has been developed to examine whether people in the model regions can imagine, or even wish, to live a corresponding life and what opportunities they have to do so. The definition of sustainable quality of life (SQoL) and the results obtained were intended to contribute to the research discussion on quality of life and sustainability and to serve sustainable development. This data science project is part of the research project “Sustainable Quality of Live” currently conducted at the Centre for Development and Environment at the university of Bern.

The overall objective of the research project “Sustainable Quality of Live” is to identify opportunities to link quality of life and sustainability. For this purpose, a written survey has been conducted in three nature parks (UNESCO Biospäre Entlebuch, Naturpark Gantrisch, Jurapark Aargau) and one control region.

In this data science project, the aim is to provide a first exploratory data analysis and to identify factors impacting life satisfaction of people. More specifically the objectives are:

1. Preprocess the collected data
2. Describe the data
3. Identify correlations and factors impacting life satisfaction

METHODS

Data acquisition

In order to collect the required data, paper questionnaires have been sent to 13,314 people. The questionnaires have been answered and returned via mail or filled-in online. A written reminder has been sent to all household a few weeks later.

The paper questionnaires were digitalized using the text recognition software Remark Office OMR and via manual data entering. The digitalized data was then merged with the results from the online questionnaire.

A preliminary cleaning was conducted in order to identify potential duplicates and unfinished questionnaires. The overall rate of return was 25%. This means from a total of 13,314 sent questionnaires 3,358 valid questionnaires were returned and considered valid.

Tools

Processing hardware: A MacBook Pro 2018 (2.2GHz Intel corei7, 32 GB memory, 1TB storage) was used for computation. It is assumed that this will be largely sufficient, as processing power and storage space required for the planned data analysis are considered to be modest.

Storage: All data, the scripts to analyze the data, the analysis results and the documentation were stored in the cloud using Microsoft's OneDrive service. This allows to have access to the data from different workstations (stationary PC and laptop). Privacy concerns, that are commonly linked to such cloud storage solutions, were considered to be negligible as the used dataset only contains anonymized data and the identification of single persons is not possible. An external hard disk is used to store incremental backups of the entire project folder. In addition, all potential code is being mirrored on GitHub.

Software: Rstudio is used as text editor and development environment allowing to create notebooks or simple scripting files. The data is being analyzed using R (version 3.6.1) and some of its many available libraries. The main libraries for data processing where "dpylr" and "tidyr". Plots where mostly created using the "ggplot2" package. Versioning of the data analysis scripts was done using "git".

Environment: In order to avoid broken dependencies (i.e. to guarantee reproducibility), we used the "renv" package that installs all used libraries locally in the project folder. Compared to alternative package manager (e.g. Conda) it does not store the version of R. However, I chose it as it stores all dependencies directly in the project folder (I still need to figure out how to do this with Conda).

Data analysis

I will use simple hypothesis testing in order to compare difference between different groups. In addition, a first regression will be applied to analyze the factors impacting life satisfaction.

DATA

Data description

The original dataset contains 147 variables and 3,358 items (questionnaires). A complete overview of the data can be found in the code book. Additional variables have been and will be calculated during the data analysis phase. Figure 1 gives an overview of the data type of each variable and the missing values.

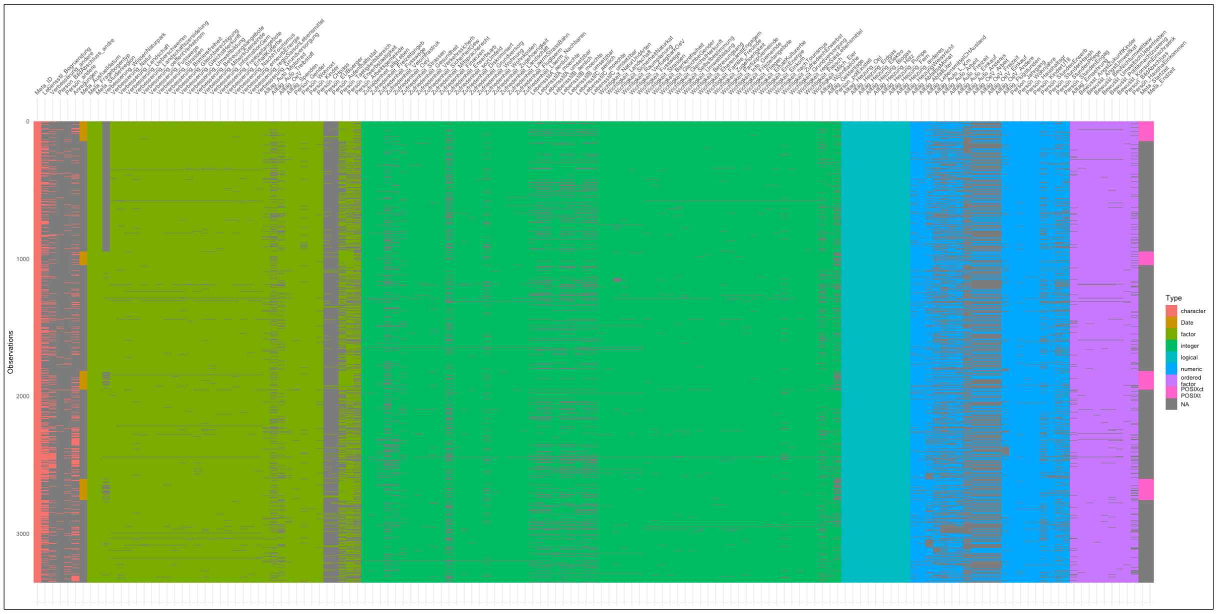
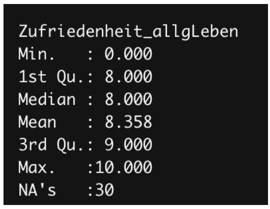
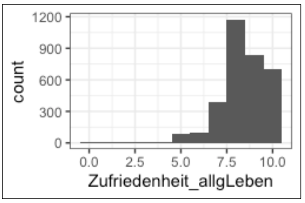


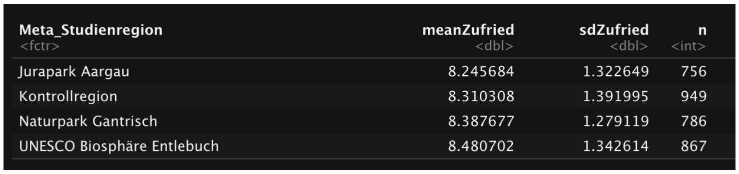
Figure 1: Data type of variables and missing values

A first overview on the overall life satisfaction is given in the Box 1. On a scale of 0 (very low) to 10 (very high) the mean satisfaction is on 8.3. A first short analysis indicates that there might be differences in life satisfaction between different groups (e.g. Parents vs. non-parents, nationality, age, income). However, it still has to be tested whether these differences are significant.

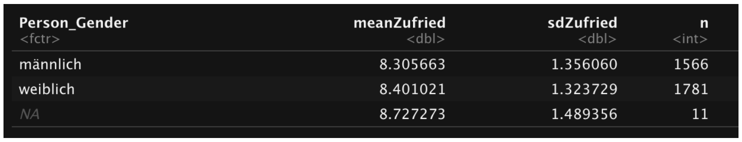


```
Zufriedenheit_allgLeben
Min.   : 0.000
1st Qu.: 8.000
Median : 8.000
Mean   : 8.358
3rd Qu.: 9.000
Max.   :10.000
NA's   :30
```

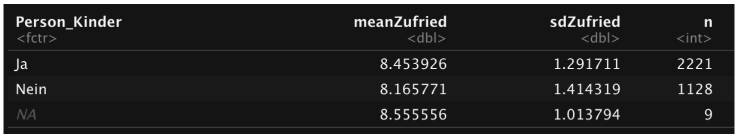




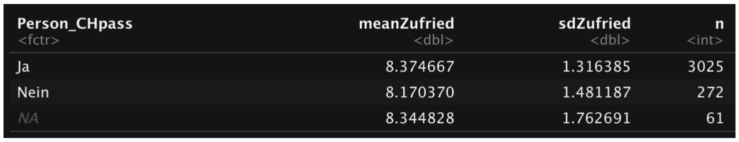
Meta_Studienregion <fctr>	meanZufried <dbl>	sdZufried <dbl>	n <int>
Jurapark Aargau	8.245684	1.322649	756
Kontrollregion	8.310308	1.391995	949
Naturpark Gantrisch	8.387677	1.279119	786
UNESCO Biosphäre Entlebuch	8.480702	1.342614	867



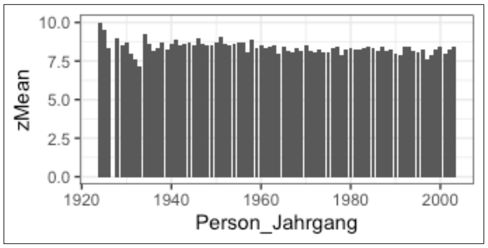
Person_Gender <fctr>	meanZufried <dbl>	sdZufried <dbl>	n <int>
männlich	8.305663	1.356060	1566
weiblich	8.401021	1.323729	1781
NA	8.727273	1.489356	11

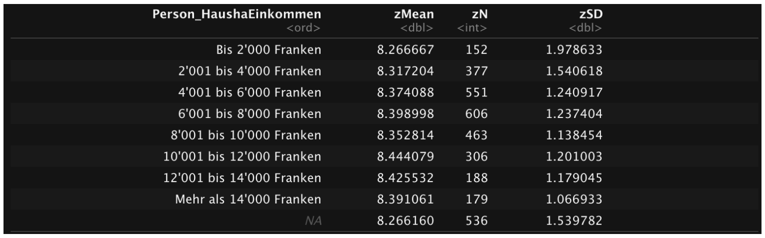


Person_Kinder <fctr>	meanZufried <dbl>	sdZufried <dbl>	n <int>
Ja	8.453926	1.291711	2221
Nein	8.165771	1.414319	1128
NA	8.555556	1.013794	9

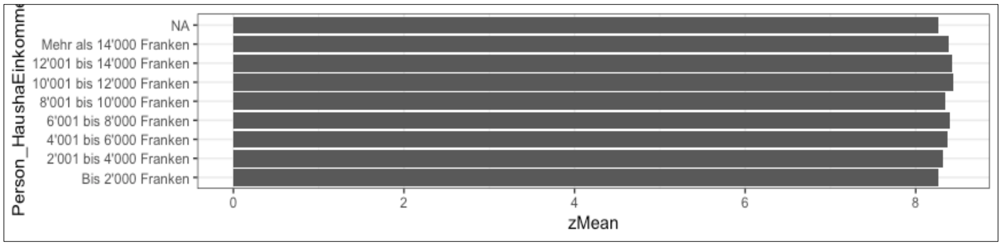


Person_CHpass <fctr>	meanZufried <dbl>	sdZufried <dbl>	n <int>
Ja	8.374667	1.316385	3025
Nein	8.170370	1.481187	272
NA	8.344828	1.762691	61





Person_HaushaEinkommen <ord>	zMean <dbl>	zN <int>	zSD <dbl>
Bis 2'000 Franken	8.266667	152	1.978633
2'001 bis 4'000 Franken	8.317204	377	1.540618
4'001 bis 6'000 Franken	8.374088	551	1.240917
6'001 bis 8'000 Franken	8.398998	606	1.237404
8'001 bis 10'000 Franken	8.352814	463	1.138454
10'001 bis 12'000 Franken	8.444079	306	1.201003
12'001 bis 14'000 Franken	8.425532	188	1.179045
Mehr als 14'000 Franken	8.391061	179	1.066933
NA	8.266160	536	1.539782



Box 1: Overview of relevant variables

Security issues

The entire dataset is anonymized, i.e. no names and addresses are stored together with the questionnaire. In theory, individuals could be identifies based on their socio-economic characteristics (e.g. age, sex, and domicile). While I consider the chances for such an identification very low, I might consider removing sensitive variables (e.g. political view, household income, …) before sharing the dataset.

METADATA

The metadata was directly stored in the RDS file (native R data format), which contains also the preprocessed data. The metadata contains the description of the data set itself, but also the description of each variable.

In addition, I used the "codebook" library to automatically create a codebook in HTML format containing a description of all metadata. This codebook is stored in the project folder together with the dataset and can be found on GitHub (https://github.com/bushroot/2020_ADS_M1/blob/master/03_codebook.html). ¨

DATA QUALITY

First checks of the data suggest a sufficiently high quality of the data. With a rate 25% the return was higher than expected (10% to 20%). This allows will allow us to provide representative results in each study area.

Overall, I consider the completeness of the data as satisfactory. Figure 2 provides an overview of the missing values for each variable. Overall, 14% of all values are missing values. These missing values are distributed that unevenly. In fact, most missing values concern question that were facultative or that follow a specific condition and therefore had not been selected. In most relevant questions, the share of missing values is below 2%. However, in a few relevant cases the share of missing values reaches up to 20%. This can be explained the difficult of the respondents to choose the appropriate categegory from the possible answers (e.g. working sector) or because of delicacy of the question (e.g. income).

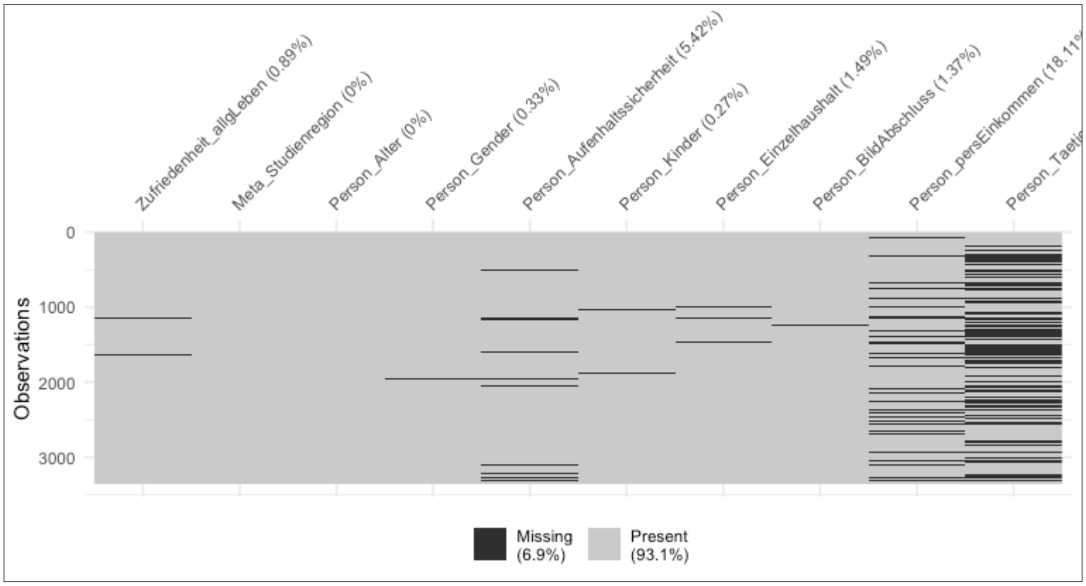


Figure 2: Missing data per variable

The number of missing values will be carefully examined for each specific analysis. Furthermore, additional plausibility and consistency checks will be conducted continuously and according to the specific analysis steps.

DATA FLOW

Figure 3 gives an overview of the dataflow in this data science project. The flow can be sub-divided in four main steps: data collection, preprocessing, analysis, and publication. The following paragraph provides more details for each step.

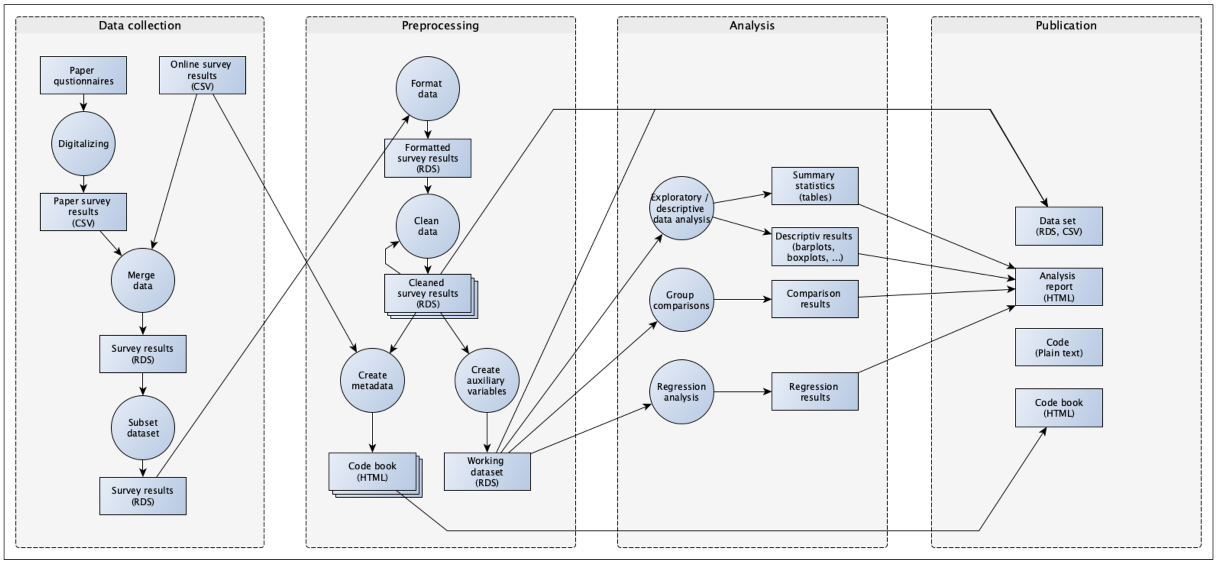


Figure 3: Data flow diagram

Data collection

In order to collect the required data questionnaires have been sent to households in 13,314. The questionnaires have been answered and returned via mail or filled-in online. A written reminder has been sent to all household a few weeks later. The paper questionnaires were digitalized using the text recognition software Remark Office OMR and via manual data entering. The digitalized data was then merged with the results from the online questionnaire. Finally, the data was reduced to a subset of variables considered relevant for this data science project.

Data preprocessing

Data formatting: The data is being first casted into the right format (continuous, ordinal, categorical) as well as checked and cleaned for the predefined data categories and ranges (e.g. ranges for continuous data, categories for categorial data).

Data cleaning: The data cleaning step includes notably checks for unrealistic outliers. Unrealistic outliers are set to NA. I assume that not all data anomalies and errors will be identified during the first data clearing. Therefore, additional data cleaning and correction will be conducted during the data analysis process. Each time a new data cleaning is conducted the data set is being stored using a different version name (e.g. data_v01, data_v02).

Metadata creation: For each cleaned dataset (survey results) a codebook will be created.

Creation of auxiliary variables: Based on the variables directly obtained in the survey, auxiliary variables will be created in order to allow further data analysis. These calculations include for instance: (1) normalization of income from household income and household size to the individual, (2) creation of binary variable of whether the person live in a single household, or (3) creation of binary variable of whether the person has kids.

Data analysis

Descriptive analysis: The descriptive analysis contains summary statistics and visuals overviews for each variable in the working dataset.

Group comparisons: This analysis step investigates the differences of life satisfaction between different groups.

Regression: A first regression analysis will be conducted in order to estimate life satisfaction based on given variables.

Data publication

The expected outputs are the analysis report (html file), the code (plain text files), the datasets (RDS, CSV), and the codebook (html). This should allow full reproducibility. However, since the data and analysis is part of an ongoing research project, the outputs will not yet be made publicly available. The code, codebook and reports will be available on GitHub. The data, however, will be keep private until the data has been published.

DATA MODELS

This project contains only one single dataset containing the results of survey. Therefore, conceptual, logical and physical models are not pertinent. The specific definition of each data attribute (variable) can be found in the code book (https://github.com/bushroot/2020_ADS_M1/blob/master/03_codebook.html).

RISKS

The following potential risks were identified:

- **Data loss:** There is a potential risk of data loss due storage devise failures or unintended data deletion. Precautionary measures include distributed storage and frequent backups. For this purpose, the project is stored on a cloud-based storage service (OneDrive) and versioned backups made on an external (encrypted) hard disk. In addition, all code is versioned using Git.
- **Unauthorized access:** There always is a minimal potential risk of unauthorized data access. Such an unauthorized access could result in the leakage of confidential data or in the loss of the data (e. g. encryption by ransomware). I am aware that the cloud-based storage service (OneDrive) causes a significant risk of unauthorized access. For this reason, only anonymized data is being stored in the cloud. To avoid the potential data loss backups were made frequently (cf. above).
- **Data collection errors:** There are two main risks of related to data collection errors. First, we did not use personalized questionnaires with unique identifiers in order to guarantee anonymity and to simplify the data collection. As a result, the questionnaires could be filled by identical people multiple. However, I consider the

risk to be negligible. Second, the paper questionnaires have been digitalized using a OGR software (Remark Office) and by manually entering the data (research assistants). I both cases errors can occur. Systematic errors are expected to be identified during the data cleaning step.

- **Pool data quality**: There is a risk of poor data quality considering the rate of return (of the questionnaires) or the number of missing values. A low rate of return increases the risk of biased survey results. Such a bias is theoretical possible since even a rate of return above 20% can be considered as high for written questionnaires. I will simply keep in mind this possible bias during the result interpretation phase. The missing values will be handled for each analysis step specifically.
- **Data analysis mistakes:** Potential data analysis mistakes will be most likely due to the incorrect application of data analysis / statistical methods and due to the incorrect interpretation of the results. In order to minimize this risk, I will search for support and advice from colleagues and peers.
- **Violation of anonymity:** The entire dataset is anonymized, i.e. no names and addresses are stored together with the questionnaire. However, individuals could be identifies based on their socio-economic characteristics (e.g. age, sex, and domicile). While I consider the chances for such an identification very low, I might consider removing sensitive variables (e.g. political view, household income, …) before sharing the dataset.

PRELIMINARY STUDIES

Life satisfaction

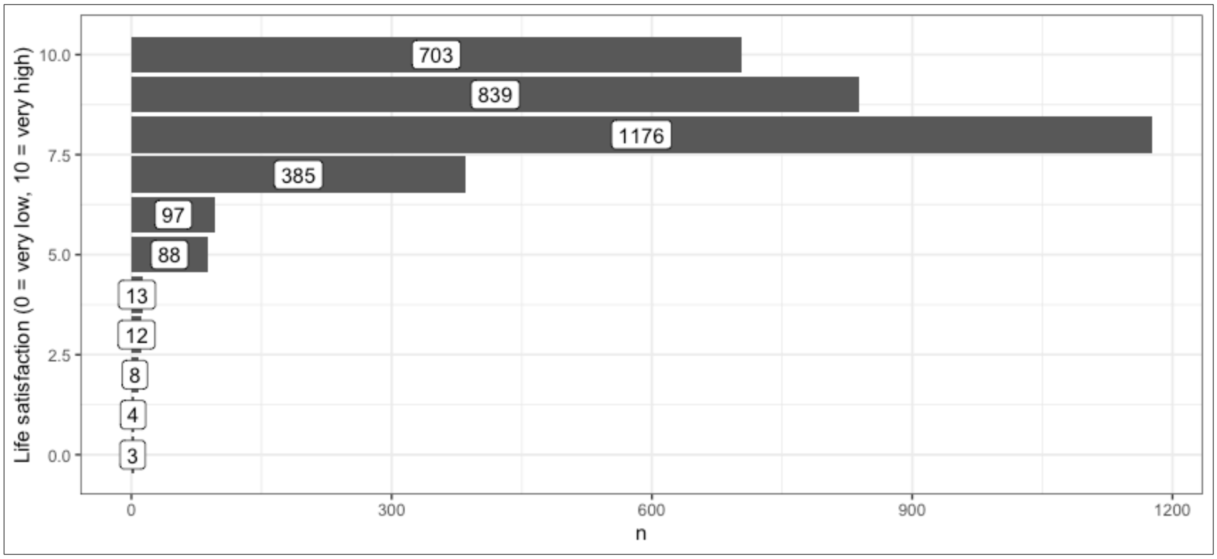


Figure 4: Life satisfaction (0 = very low; 10 = very high)

Shapiro-Wilk normality test: W = 0.8598, p-value < 0.00000000000000022
The p-value is very small. We can reject the 0 hypothesis that the data has a normal distribution.

Regions

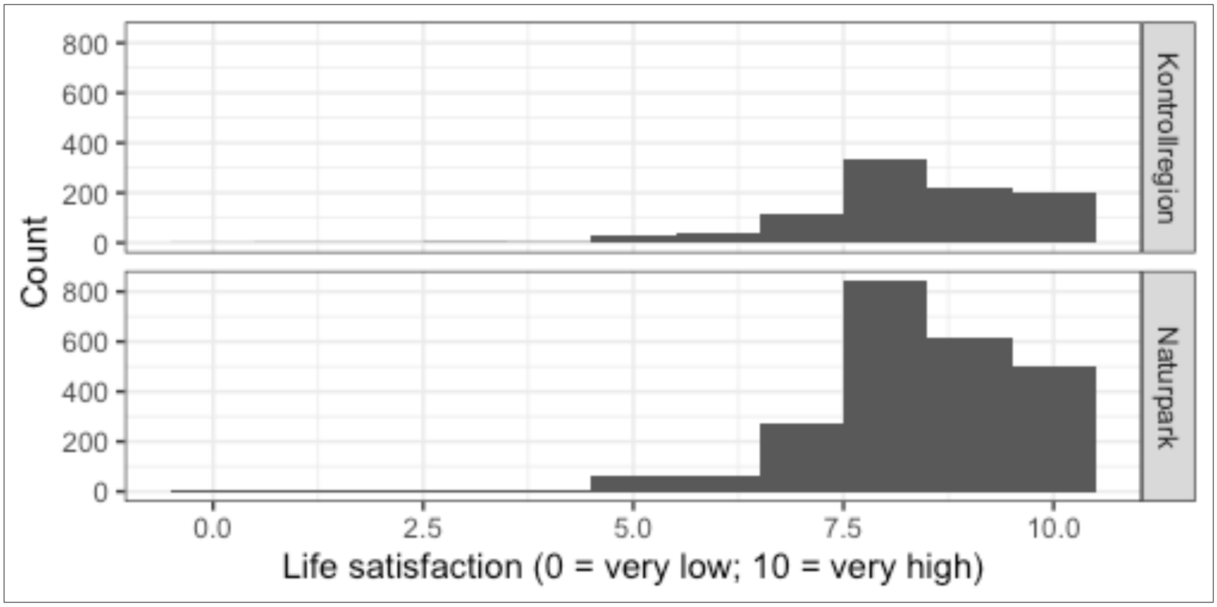


Figure 5: Life satisfaction by region

Mann-Whitney U test: Life satisfaction by Region
Results: W = 1096739, p-value = 0.2735
There is no significant difference in life satisfaction between the study areas.

Sex

Table 1: Life satisfaction by sex

Sex	**Mean**	**SD**
Male	8.305	1.356
Female	8.401	1.324

Mann-Whitney U test: Life satisfaction by Sex
Results: W = 1314094, p-value = 0.03555
Women have a significantly higher life satisfaction then men.

Children

Table 2: Life satisfaction by children

Childern	**Mean**	**SD**
Yes	8.454	1.291
No	8.166	1.414

Mann-Whitney U test: Life satisfaction by Children
Results: W = 1374798, p-value = 0.000000007068
Parents have a significantly higher life satisfaction then non-parents.

Single household

Table 3: Life satisfaction by household size

Singel hh	**Mean**	**SD**
Yes	8.079	1.711
No	8.398	1.268

Mann-Whitney U test: Life satisfaction by Single household
Results: W = 561254, p-value = 0.002834
People in living in non-single households have a significantly higher life satisfaction people living in single households.

Age

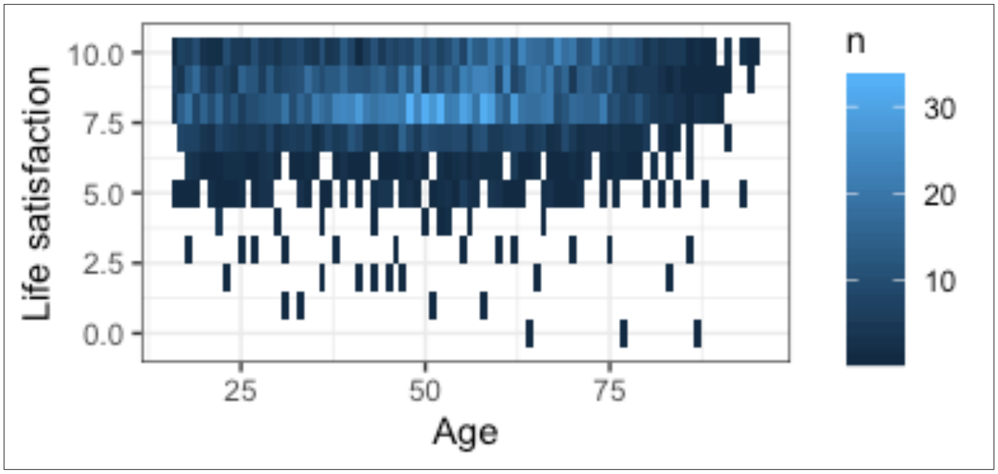


Figure 6: Life satisfaction by age

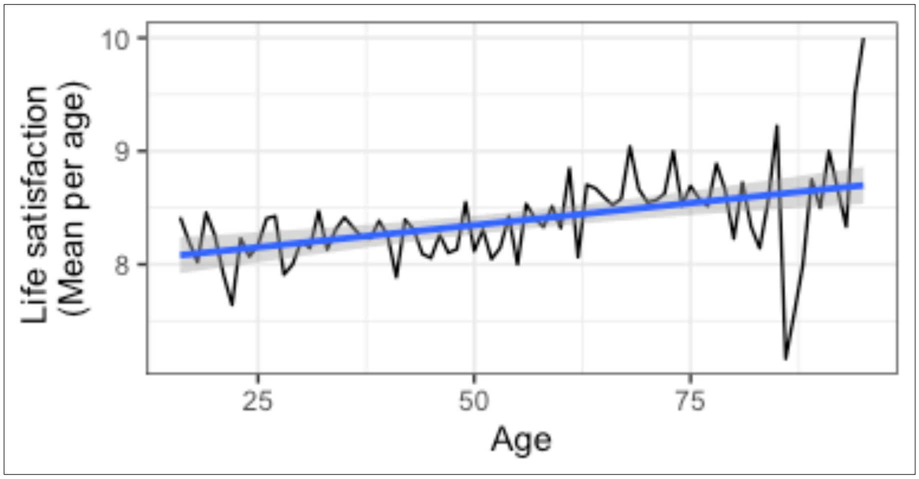


Figure 7: Mean life satisfaction by age

Kruskal-Wallis rank sum test: Life satisfaction by Age

Results: Kruskal-Wallis chi-squared = 176.93, df = 78, p-value = 0.000000001192

People with different ages have a significantly different life satisfaction.

Income

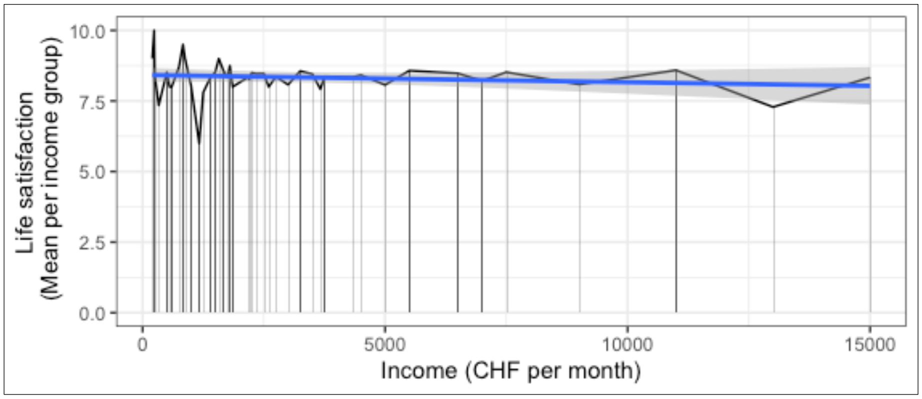


Figure 8: Life satisfaction by mean income

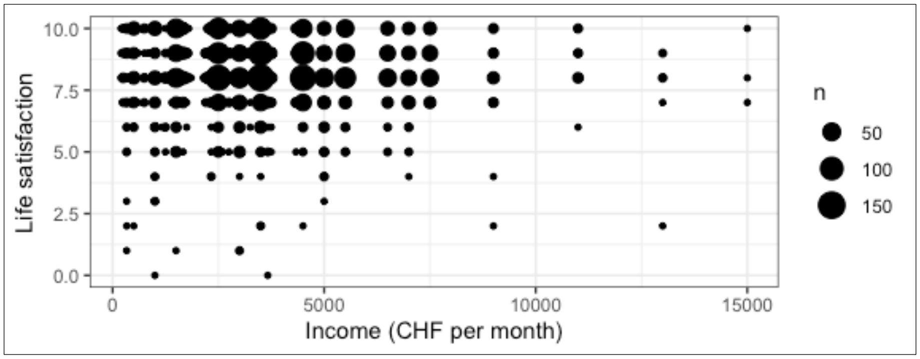


Figure 9: Life satisfaction by income

Kruskal-Wallis rank sum test: Life satisfaction by Income

Kruskal-Wallis chi-squared = 56.849, df = 42, p-value = 0.06278People with different incomes have not a significantly different life satisfaction

First attempt of linear regression

```
Call:
lm(formula = Life satisfaction ~ Age +  Sex + Children + Single hh + income, data = d)

Residuals:
Min         1Q          Median    3Q    Max
-8.5130    -0.5482    -0.0887    0.8346     2.4290

Coefficients:
                    Estimate          Std. Error        t value      Pr(>|t|)
(Intercept)    7.25547267     0.15907829     45.609      < 0.0000000000000002 ***
Age               0.01142662     0.00170323     6.709        0.0000000000238 ***
Sex               0.13212334     0.04933400     2.678        0.00745 **
Childden       0.03320145     0.06185513     -0.537       0.59148
Single hh       0.42078795     0.07660454     5.493        0.0000000431854 ***
Income          0.00002874     0.00001241     2.316        0.02062 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error:
1.277 on 2713 degrees of freedom (639 observations deleted due to missingness)
Multiple R-squared: 0.03188, Adjusted R-squared: 0.03009
F-statistic: 17.87 on 5 and 2713 DF, p-value: < 0.00000000000000022
```

The age, gender and household composition seem to have an significant impact on life satisfaction. However, the overall fit of the model is very low (R squared = 0.03009). I.e. the model explains very little of the variability.

Possible improvement: check better for assumptions, e.g. multicollinearity, normal distribution of error terms and homoscedasticity

CONCLUSION

- Life satisfaction is different according to different groups. For instance, significant differences among the population can be found according to the sex (female higher), whether people have children or not (higher with children), whether people live in a single household (higher if not) and according to their age (higher if older).
- The regression result do not (yet) provide good results and needs substantial improvement.

- Further interesting analysis include the inclusion of additional variables and the statistical inference to the entire population.