PDEng Program
Data Science

Visualization of GGD Screening Youth Data

Prepared by

Abel Gebresilassie

Sarah Ibrahimi

Valentine Tuyishime

Prepared for

GGD West-Brabant and GGD Hart voor Brabant

February 2017

Visualization of GGD Screening Youth Data

Abel Gebresilassie

Sarah Ibrahimi

Valentine Tuyishime

PDEng Program Data Science

Jheronimus Academy of Data Science

Eindhoven University of Technology, the Netherlands

Convent Mariënburg

Sint Janssingel 92
5211DA's-Hertogenbosch

University supervisors: Jack van Wijk

Industrial supervisors: Ike Kroesbergen

Elly de Boer

February 2017

Executive Summary

The mission of the GGDs across the Netherlands is to guard, protect and promote public health within the municipalities associated to each of the GGDs. The GGDs fulfill their mission through monitoring, advising community, and assisting in decision-making processes at municipality level. The GGDs focus on the well-being of the whole population: children, adults, and elders. It is in this regard that GGD conducts research in the category of school age students in order to get insights in their daily life and the challenges they face.

The GGD West-Brabant and the GGD Hart voor Brabant collected data from 13 to 15 year old pupils. This collection has been done using questionnaires administered to respondents from all municipalities in both regions. Using the collected data for three consecutive school years (2013/2014 to 2015/2016), the GGDs are interested in finding patterns in the data, and developing an effective way to visualize the data.

This report presents the results of our analysis and describes the functionalities of the developed interactive visualization tool. The tool supports pattern identification and benchmarking whereby the user is enables to compare municipalities with respect to different variables in both regions. It also supports a trend analysis of lifestyle related variables for individual municipalities in the Hart voor Brabant region. With the tool, changes of identified variables over time for each municipality compared to the weighted average of the region can be accessed easily.

This tool is a prototype that can be adapted and scaled to enable the analysis of other similar datasets.

Table of Contents

Executive Summary

1 Management Introduction 1

1.1 Problem background 1

1.2 Goal specification and added value 1

1.3 Strategy 2

1.4 Results 3

1.5 Conclusions and recommendations 3

2 Data value chain 4

2.1 Dataset information 4

2.2 Data Science approach 5

2.3 List of data, variables, and abbreviations 8

3 Dataset characteristics and structure 9

3.1 Entity relation diagram 9

3.2 Entity type and attributes 9

3.3 Type of data 10

3.4 Scale 11

4 Results 12

4.1 Pattern Identification 12

4.2 Profile of a municipality 17

4.3 Trend Analysis 20

5 Conclusions 25

6 Recommendations 26

Appendix A 27

1 Management Introduction

1.1 Problem background

The 25 GGDs in The Netherlands form the Dutch municipal health organization for preventive healthcare. The GGDs have several focus groups, in particular, children, adults and elders. Among other tasks and responsibilities toward their focus groups, the GGDs are in charge of monitoring, advising, and informing the municipalities that they serve about preventive healthcare. In 2015, the government decentralized some public health related responsibilities to the municipalities. To ensure the success of these reforms, GGDs across the country have been assigned additional responsibilities to inform public health related decisions at municipality level.

It is in this regard that the GGD West-Brabant(WB) and the GGD Hart voor Brabant(HvB) conduct studies and collect data on different public health aspects concerning their focus groups. Using data collected from past years, the GGDs are interested in finding relationships and patterns in the data as well as in creating simple and useful means to present data to their different users, namely: GGD researchers, municipality policy makers, and policy advisors.

1.2 Goal specification and added value

For their research, the GGDs, WB and HvB, collect data from samples of their focus groups. To collect this data, they use questionnaires that are customized to different age groups. In this project, our team of three PDEng Data Science trainees focuses on data collected from school children who are between 13 and 15 years old. The data covers three consecutive school years: 2013-2014, 2014-2015, and 2015-2016. The data was collected using questionnaires administered to pupils from schools located in the regions of West-Brabant and Hart voor Brabant. The questions concern the following five categories:

1. Personal identification
2. Physical health
3. Psycho-social problems
4. Social life

5. Socio- Economic life.

In order to extract knowledge from the collected data and add value to the main task of the GGDs, our team set the following goals:

1. Explore the data and identify patterns; and
2. Design an interactive visualization tool that enables the user to get insights from the data, and to compare different areas (within the region and across the region).

These goals are in line with the goal and mission of the GGDs, since they focus on promoting a healthy lifestyle through their programs and advices that they provide to local institutions. Getting insights into the data through an exploratory analysis and having a visualization tool will enable the GGDs to:

1. Assess the performance of municipalities for different adolescent related features
2. Identify problems among school age youth
3. Facilitate the communication between GGD and other stakeholders, especially policy makers and policy advisors

1.3 Strategy

To achieve our goal, we addressed the project from two complementary perspectives:

1. Execution of an exploratory analysis of data (pattern identification, trend analysis, and municipality ranking);
2. Design of an interactive visualization tool that supports the results of our analysis and enables benchmarking.

Concerning the exploratory analysis, we applied correlation analysis techniques. We also performed a trend analysis of lifestyle related variables for the Hart voor Brabant municipalities.

We used R Shiny to develop a simple but useful interactive visualization tool for both GGDs. The targeted end-users are GGD researchers with some technical skills. However, thanks to its simplicity, the tool can also be used to facilitate the communication between GGD experts and municipality policy and decision makers.

1.4 Results

For the correlation analysis, ranking and benchmarking of performance, we focused on the data collected in school year 2015-2016 from West-Brabant and Hart voor Brabant. We aggregated data up to municipality level because the proposed users act on this level.

Correlation analysis was done by computing the correlation between pairs of variables for each region and by visualization. For example, we observed a positive correlation in West-Brabant between risk of psychosocial problems and smoking for these high school students. The visualization of the association between these two variables shows that municipality Werkendam has both a high occurrence of health problems and risk of psychosocial problems in 2015-2016.

A similar analysis can be done using the interactive interface that we developed. It enables the user to select his interest, and make a selection by municipality to get the performance of that municipality on different variables or a selection by variable to see the performance of different municipalities on that specific variable. The user has also the option to select which region and / or which variables category to focus on.

Our trend analysis focused on variables in the lifestyle category in one of the two regions, Hart voor Brabant. The aim was to assess how those variables evolved over the last three years. We visualize the trends by using bar charts. The user can select a variable or municipality to see the trend of the selection compared to the average of the region. This visualization tool is important for policy makers since it quickly shows the performance of the municipality compared to the whole region and the trend overtime.

1.5 Conclusions and recommendations

The developed visualization tool provides three complementary views: the municipality performance, the municipality profile and the trend analysis. Using this tool, the user can see and benchmark the performance of municipalities with respect to different variables, create an individual municipality profile with respect to all the variables, as well as follow variable changes over time and compare their performance to the region-weighted average.

The development of the tool is in a prototype phase; it needs improvements and adaptions to ensure scalability to other datasets.

2 Data value chain

Every year the GGDs send questionnaires to all the 2nd year students in high schools. Each GGD does this for its own region. The students fill in the questionnaires at school and the GGD receives the questionnaire and transform them into tabular data.

The datasets contain information about 2nd year high school students, who attend school in one of the regions and live in the regions of the GGDs West-Brabant and Hart voor Brabant. The information obtained can be divided in four categories: Health, Lifestyle, Socio-economic factors and Psychosocial behavior. The purpose of our project is to provide a benchmark tool that gives the GGDs the opportunity to explore the data and to perform a benchmark between the different municipalities. Until now, the GGD only explored the data for each municipality individually at district level and presented the distributions of answers of questions in a report. However, the GGD is also interested in relations between variables, trends over time and comparisons between municipalities.

For this analysis, we received six datasets:

- GGD Hart voor Brabant 2013-2014
- GGD Hart voor Brabant 2014-2015
- GGD Hart voor Brabant 2015-2016
- GGD West-Brabant 2013-2014
- GGD West-Brabant 2014-2015
- GGD West-Brabant 2015-2016

2.1 Dataset information

The business case

The GGD is an organization that performs research on several health related domains. Questionnaire data from students is used to inform both schools and municipalities about the trends they see in the data. For the school this is mainly informative, but at the municipality level, policy makers can use the results to change their policy.

The current data analysis by the GGD is on a descriptive level. However, there is much more information in the data that is extracted at the moment. Such information can produce added value

to the policy makers, since the results will give them additional insights in the behavior and needs of the students living in their municipality.

The company's questions and objectives

The GGDs WB and HvB described a desired picture of the analysis and presentation of the results. Their picture consists of the following parts:

1. Correlations between indicators

   For now, the GGDs analyze all questions/indicators separately, and do not study relations between indicators. Researchers at the GGDs are interested in discovering correlations between indicators, since this can reveal new insights.

2. Pattern recognition

   The GGD is interested in patterns in the data, for example, clusters of students with the same behavior.

3. More attractive ways of visualization

   The GGD is currently presenting 'static' reports in pdf or printed format. They would like to have an interactive user-dynamic tool that invites use and can be tuned easily. Furthermore, it should be easy to adapt to new datasets.

4. Benchmark

   The GGD would like to have a benchmark functionality in the tool to compare municipalities and districts. Out of all the objectives, the GGD considered the benchmark functionality to be the most important one. In this way, GGD researchers who will use the tool will be able to give recommendations to policy makers.

2.2 Data Science approach

The six datasets had the same format containing data from questionnaires. Each row in the dataset represents a student, each column represents a question from the questionnaire or a score deducted from one or more questions.

Tasks

Data Cleaning

Since the questions of the dataset changed over the years, combining all the datasets is not straightforward. Since only a few questions of the datasets were asked in all three years and in both regions, only a few indicators remain when combining the datasets. This will limit our analysis. For this reason, we decided to use only the datasets from 2015-2016 of both West-Brabant and Hart voor Brabant for the main analysis. For the trend analysis, we decided to combine the datasets of three years for only one region, Hart voor Brabant. The analysis for West-Brabant can be done in the same way, but with other indicators.

Before merging the datasets, we needed to make sure the scales for questions in different datasets were the same. Furthermore, we decided to code all the missing values with the same value to avoid confusion. For the analysis, we decided to include the sample with missing values in the dataset for the total count of people per municipality. However, since we were analyzing problematic behavior, we only took a subset of the data to aggregate. In this subset, only values that are considered as problematic for each variable were taken into account. Furthermore, we did not use the missing values for this part.

Exploratory Data Analysis

After cleaning the data, we did an exploratory data analysis. We calculated some descriptive statistics and looked for correlations. Since we were looking for results on municipality level and not on individual level, we decided to aggregate the data on municipality level.

For pattern analysis, we started with a cluster analysis. Due to the restricted time for the project, we had to set priorities and decided not to pursue this further. For the trend analysis, we compared data for three years of Hart voor Brabant. The focus was on indicators related to lifestyle. After looking for correlations between variables for the whole dataset, we decided that it would be more relevant for the GGD to search for a correlation on municipality level. This makes it easier to compare behavior in different municipalities.

Development of an interactive visualization tool

For the benchmark, the trend analysis, and relations between variables are already contributing to this. Furthermore, we decided to do the following:

- For each municipality, show the performance of each variable
- For each variable, show the performance of each municipality

We decided to do the analysis on municipality level both on value, and on rank.

From our results, conclusions were drawn and we presented suggestions for future work.

Strategy

The strategy was according to the diagram in Figure 2.1. The first tasks were performed as a team, and after a while, tasks were divided among team members and executed in the order as indicated below. In the meantime, the report was written and conclusions were drawn along the way.

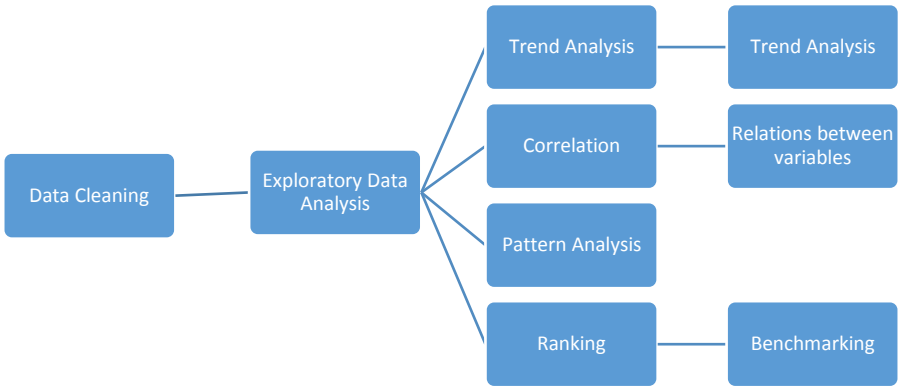

Figure 2.1. Strategy of the project; data handling, and data analysis.

Since the analysis had to be performed on the system of the GGD, and the final tool had to run on this system as well, we decided to use software that was already enrolled on their platform. The only software suitable for this type of analysis was R, so we used R and decided to create the interactive tool with Shiny. We had a weekly meeting with the customers to discuss our progress, test the visualization tool, and determine the way forward.

2.3 List of data, variables, and abbreviations

The GGD provided six datasets. We decided to continue with four of them:

- GGD Hart voor Brabant 2013-2014
- GGD Hart voor Brabant 2014-2015
- GGD Hart voor Brabant 2015-2016
- GGD West-Brabant 2015-2016

The two datasets we did not use for this project are:

- GGD West-Brabant 2013-2014
- GGD West-Brabant 2014-2015

The reason for this was the great difference between variables in the datasets. By merging all the datasets, we could not take many variables into account for our analysis. Furthermore, the data cleaning took more time than expected. For this reason, we decided to use only the data from GGD Hart voor Brabant for the trend analysis.

Details about the variables and their meaning can be found in Appendix A.

3 Dataset characteristics and structure

3.1 Entity relation diagram

The datasets in this project are collected from the answers that students of second year high school gave to the questionnaires that GGD sends out every school year. As mentioned before in Chapter 1, the data can be divided into different categories of students' profile.

3.2 Entity type and attributes

There are five main categories that can describe each student in the survey. These categories include Personal Identification (excluded in the data analysis), Health Problems, Psychosocial Problems, and Lifestyle, and Socio-Economic life of individual students. A schematic diagram that shows each category and some variables (attributes) associated to each category is given in figure 3.1.

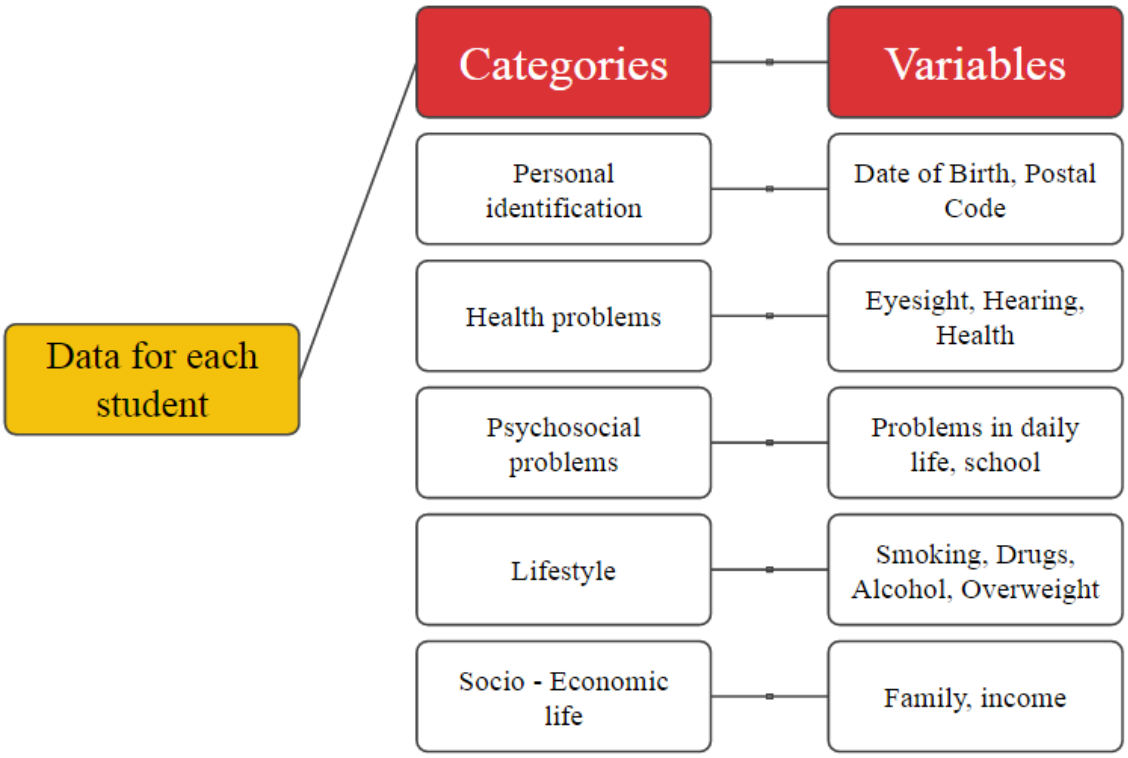

Fig 3.1. Overview of the dataset for each student and the categories and some of the variables under each category.

3.3 Type of data

The type of data is multivariate tabular data in the format of a CSV file. Except for some variables like the CBS code (Centraal Bureau voor Statistiek code), height and weight of each student, the type of the data is mostly categorical and ordinal. The data is per respondent from all the schools that are involved in the survey. These schools are located in the districts of each municipality. The sample size differs from year to year and from district to district. We got six data sets from both GGDs, but we only worked on four of them in this project. A summary of each dataset is given in table 3.1.

Table 3.1 Dataset summary with respect to the number of variables, sample size and sample rate.

Name data set	**Number of variables**	**Sample size**
GGD Hart voor Brabant 2013-2014	85	10245
GGD Hart voor Brabant 2014-2015	86	9409
GGD Hart voor Brabant 2015-2016	109	8934
GGD West Brabant 2015-2016	72	7124
GGD West Brabant 2013-2014	*80* *Not used*	*7481* *Not used*
GGD West Brabant 2014-2015	*80* *Not used*	*7521* *Not used*

3.4 Scale

As mentioned above, the type of data is mostly categorical. Some variables are binary (0,1) coded, others are multivalued, and some are also ordinal. For example, the variable "gewicht5cat" that explains the overweight condition is coded as 1 ("severely underweight"), 2 ("underweight"), 3 ("normal weight"), 4("overweight"), and 5 ("Severe obesity"). The data analysis is done only on those parts of the data that relate to problems according to GGD. Again using the above example, for GGD, 4("overweight"), and 5 ("Severe obesity") are considered as problematic cases. Therefore, in our analysis we only focused on these groups. After some discussion with GGD, we decided to use the data at municipality level. Thus at the end, the final data we used is aggregated on a municipality level. There are 18 and 27 municipalities in the West-Brabant (WB) region and the Hart voor Brabant (HvB) region, respectively.

4 Results

Figure 4 presents an overview of the tool. It consists of two parts that use different datasets: the trend analysis that uses data from three years and the part about the municipality profile and performance that uses data from 2015/2016. This second part is divided in two subparts: one view per municipality and another view per one or two variables. The view on municipality consists of a ranking part per region and for both regions, and a percentage part sorted in three different ways: alphabetically, by value, and by deviation from average. The view per municipality consists of two bar plots of two variables and a scatterplot showing the association between variables in both bar plots. In this chapter, we will further elaborate on the analysis and the functionality of the tool.

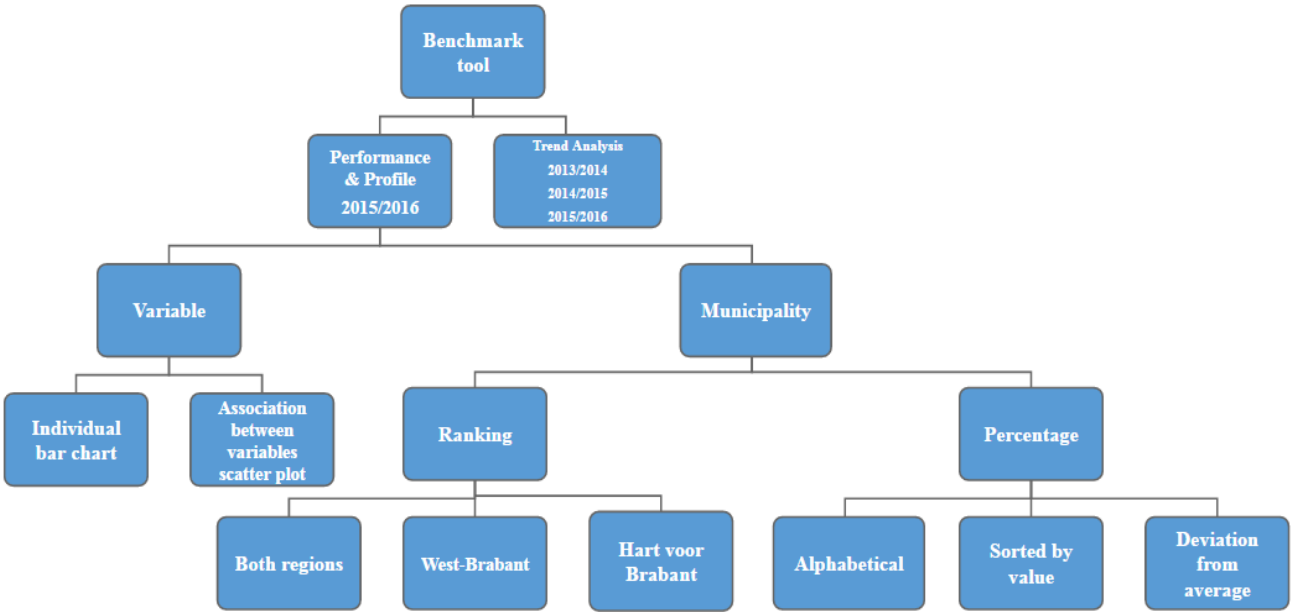

Figure 4. Overview of the interactive tool.

4.1 Pattern Identification

Correlation Matrix

The aim to identify patterns in the data was to see whether there is a link between variables, or whether certain groups are more likely to show certain attributes. To achieve this, we used correlation / association analysis. A correlation as an association between two variables does not

necessary imply that one causes the other. Both can be caused by something completely different, or it could simply be that children who show one characteristic often show the other.

We focused on data collected in the school year 2015-2016 from West-Brabant and Hart voor Brabant. The data was preprocessed to have common variables in both region for effective identification and benchmarking. To prevent multicollinearity of variables in our analysis, we reduced the number of features/variables. As some variables were indicators, or categories obtained from aggregating different other features, we kept the indicators and/or categories and removed the individual variables.

Figure 4.1 and 4.2 show the correlations matrices of different variables per regions. The correlations were calculated based on aggregated data per municipality.

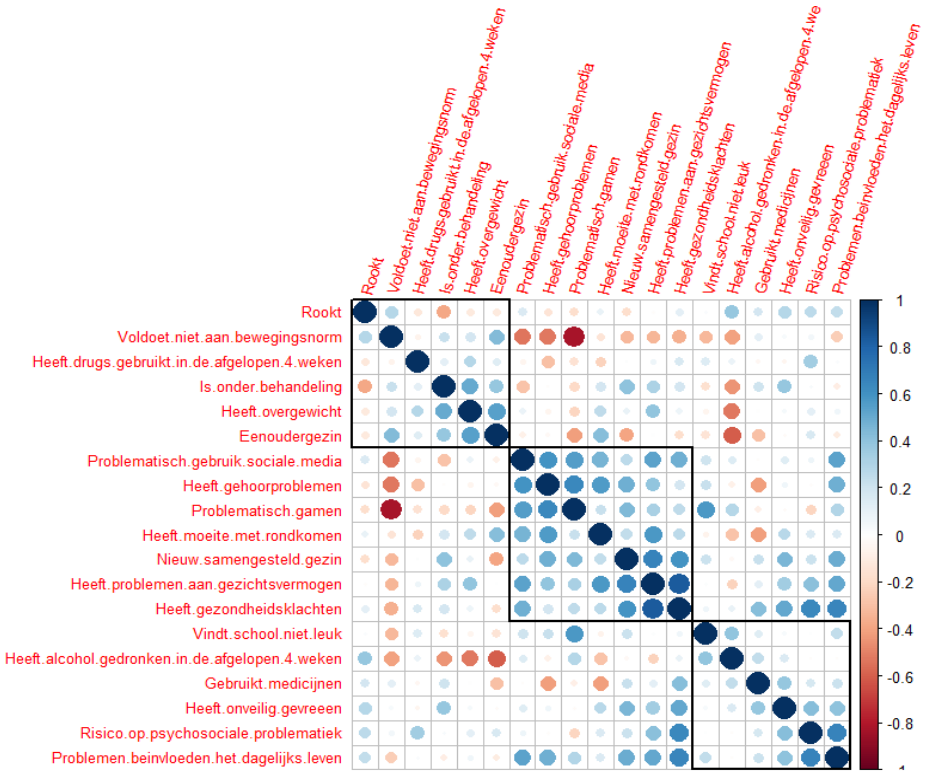

Figure 4.1.1. Correlation of variables in West-Brabant

Figure 4.1.1 shows a positive correlation of 0.6 between “Heeft gehoorproblemen” and “Problematisch gamen”. The same significant positive correlations of 0.64, 0.66 are observed

between “Risico op psychosociale problematiek” and “Problemen beïnvloeden het dagelijks leven”, and between “Heeft problemen aan gezichtsvermogen” and “Heeft gezondheidsklachten” respectively.

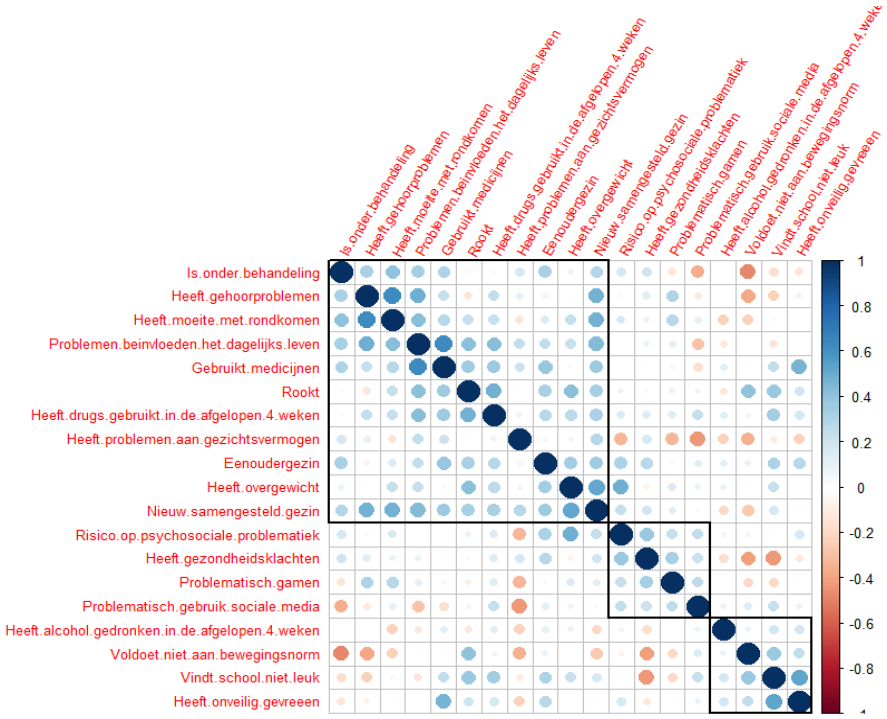

Figure 4.1.2. Correlation of variables in Hart voor Brabant

In Figure 4.1.2 of correlations in Hart voor Brabant, we observe a significant positive correlation of 0.62 between “Heeft gehoorproblemen” and “Heeft moeite met rondkomen”.

Briefly, these two figures show mixed correlations between variables in both regions that are not easy and clear to interpret. To enable interpretation and use of associations among variables, we integrated an interactive scatterplot in our interface, whereby the user can choose and zoom in to get the needed information.

Scatterplot

The interactive scatterplot developed and integrated in the interface visualizes the association between variables, and the corresponding position of municipalities, as well. We used a scatterplot of the data with two variables, one on the x- axis and another on the y-axis. The dots represent municipalities, and the size of the dots shows the number of respondents per municipality.

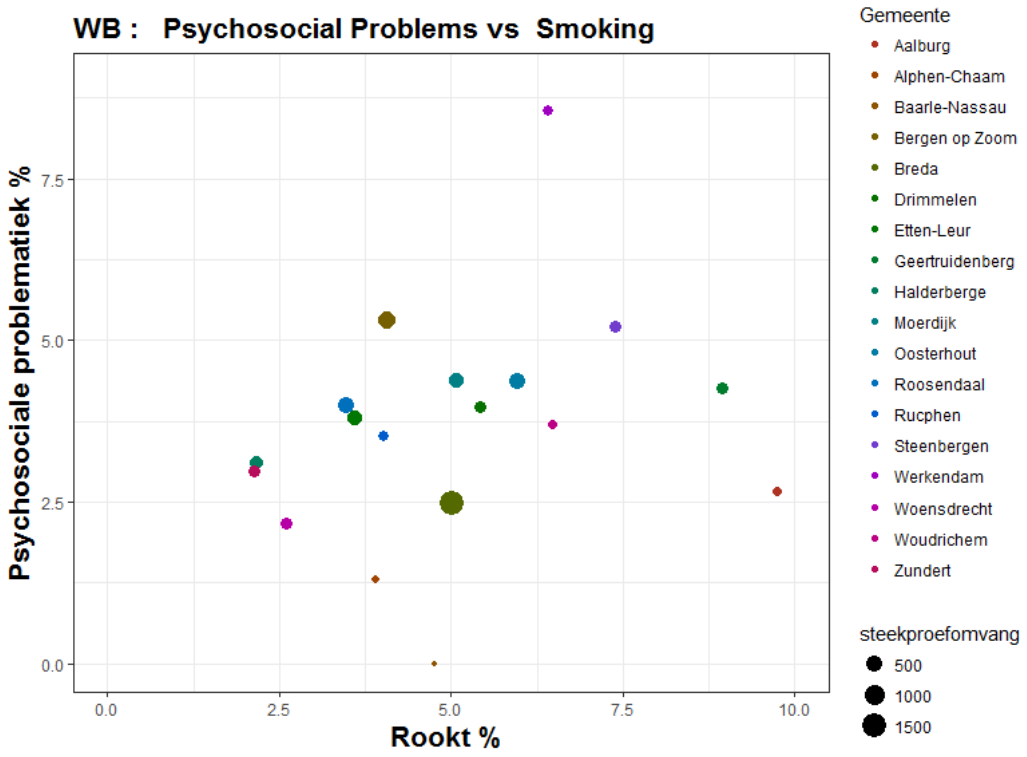

Figure 4.1.3. Psychosocial problems Vs smoking in West-Brabant

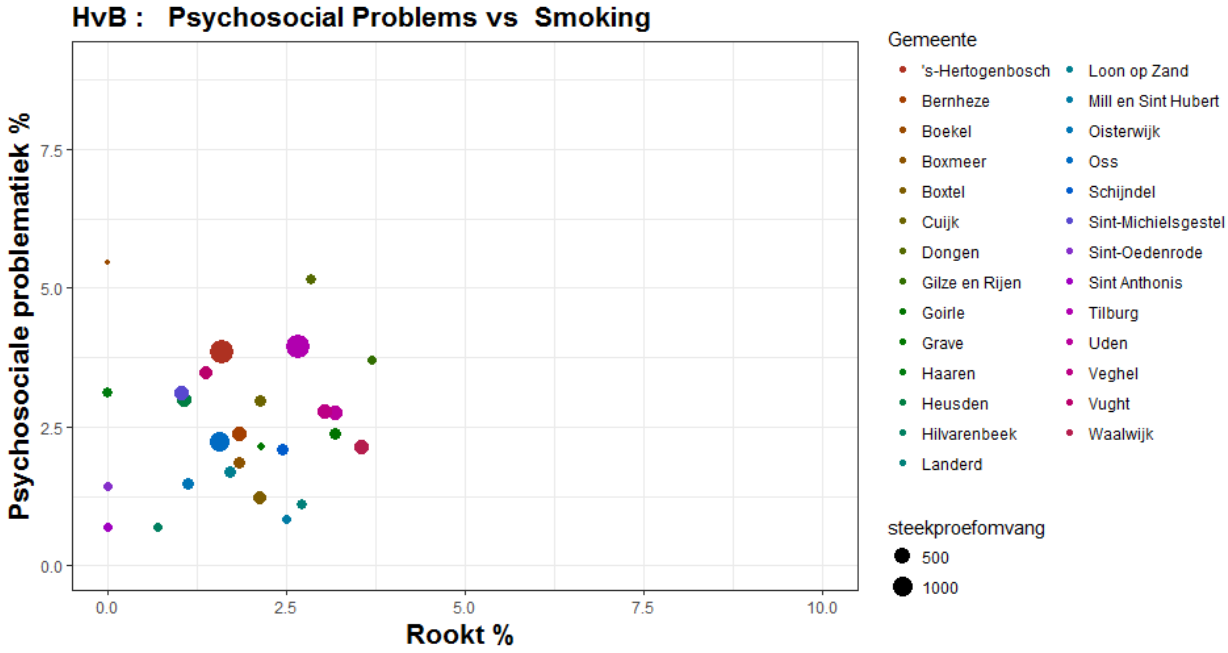

Figure 4.1.4. Psychosocial problems versus Smoking in Hart voor Brabant.

Figures 4.1.3 and 4.1.4 show the association between "risk of psychosocial problem" and "smoking" for West-Brabant and Hart voor Brabant respectively. In West-Brabant, we observe a

very low correlation of 0.27 between “risk of psychosocial problem” and “smoking”. We observe outliers such as at Werkendam with respect to psychosocial problems

In Hart voor Brabant, with respect to the same variables, we do not observe any correlation but we see outliers such as at Boekel with a high level of psychosocial problems and a low level of smoking, and Dongen with a relative high percentage with respect to both smoking and psychosocial problems.

A similar analysis of association is enabled by the interactive visual tool. The user can select variables and region of focus. The results are bar charts with the percentage levels of municipalities and scatterplots showing the relationship between selected variables. Figure 4.1.5 shows the association between “overweight” and “rate of pupils who do not meet standard level of body exercise” as well as corresponding municipalities’ levels. From both bars charts and scatterplots, we observe that the rate of pupils who do not meet standard level of body exercise is above 50% in all the municipalities in West-Brabant. The identification of such patterns can facilitate the user to make decisions that are relevant to address a specific problem.

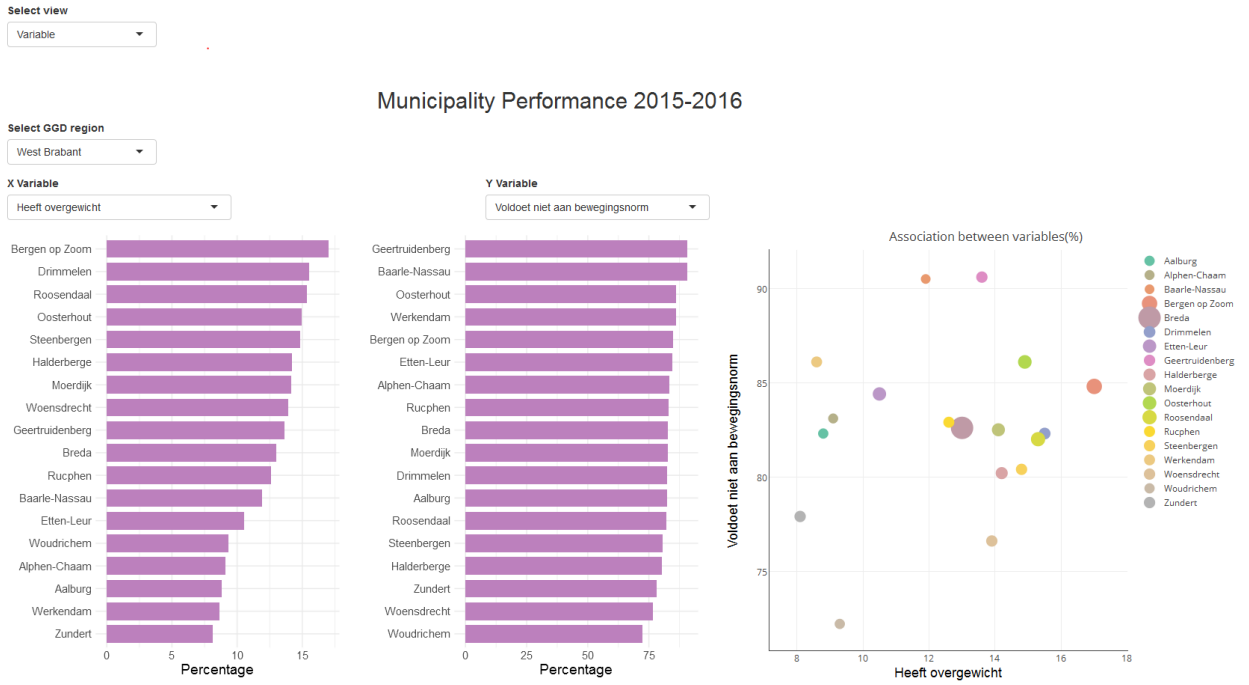

Figure 4.1.5. Image of the scatter plot embedded in the interactive visualization tool

4.2 Profile of a municipality

A desired feature was to be able to provide an overview of a municipality, such that quickly can be understood what the specific issues are. To create such a profile for each municipality, two views were defined. The first view presents the performance of a municipality on each variable based on a *rank* compared to other municipalities. In this view, the user can see how a municipality scores on a certain variable, either in comparison to the municipalities of both regions or the municipalities of its own region. The second view presents the performance of a municipality on each variable based on its *score* for each variable. To get an overall idea, the user must be enabled to switch between municipalities easily.

Ranking

In the plots below, a low ranking score indicates less problems. In Figure 4.2.1, the municipality Etten-Leur is shown compared to the other municipalities in both regions. As we see, it does not have high rankings. If we look at the municipality Baarle-Nassau compared to other municipalities in its own region, see Figure 4.2.2, we see that it has many variables on position one. This means it scores very low on problematic areas, compared to the other municipalities. Thus, a lower rank means less problems.

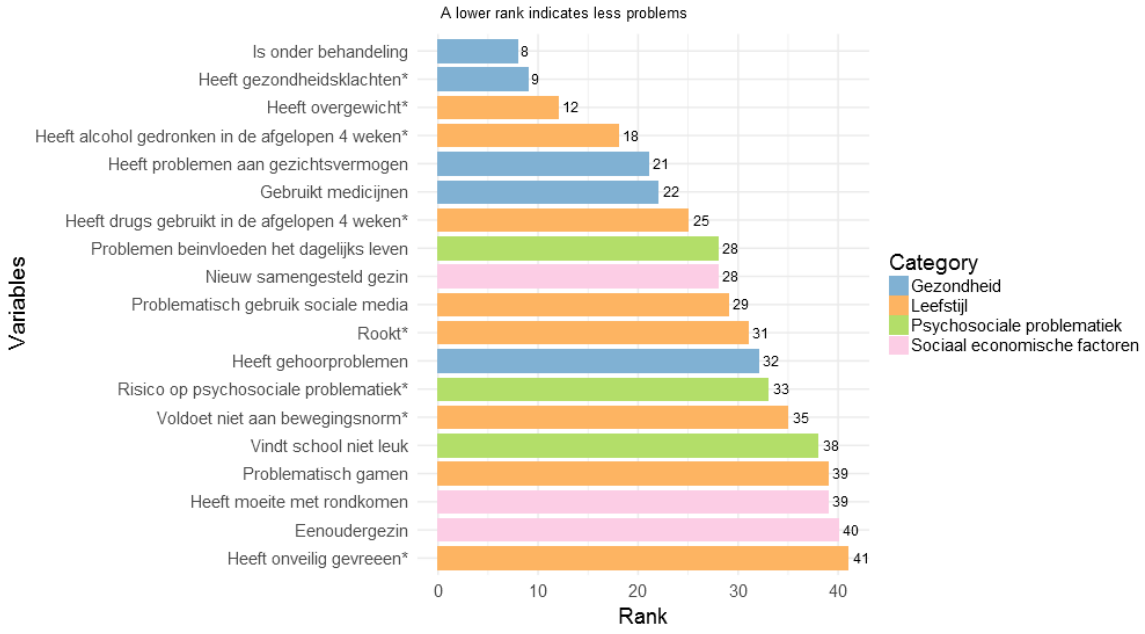

Figure 4.2.1. Ranking of variables in the municipality Dongen

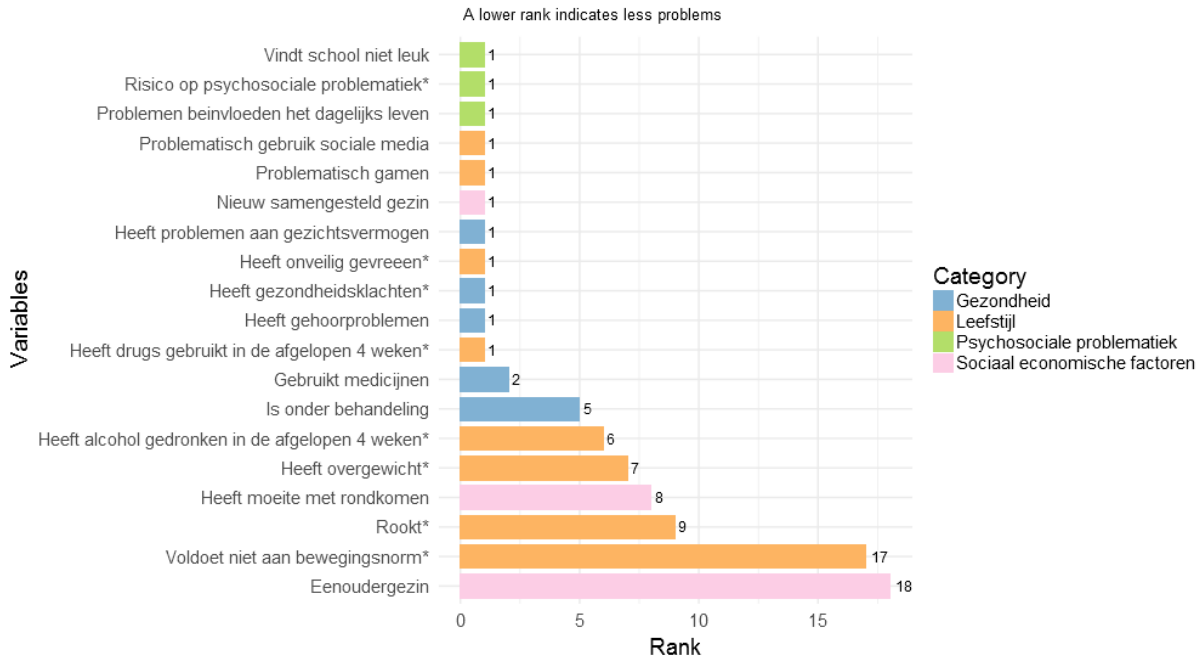

Figure 4.2.2. Ranking of variables in the municipality Baarle-Nassau

Percentage

In this view, we see the percentage of the students in the problematic area of each variable. We developed three ways to show the data. The first one is alphabetically, the second one is sorted in descending order, and the third one shows the deviation with the average value for all the municipalities in both regions. For example in Figure 4.2.3, 'Voldoet niet aan bewegingsnorm' has a high score for Baarle-Nassau; more than 80% of the respondents do not exercise enough according to GGD's standards. However, the average, which is represented as a grey bar in the same view, is high as well.

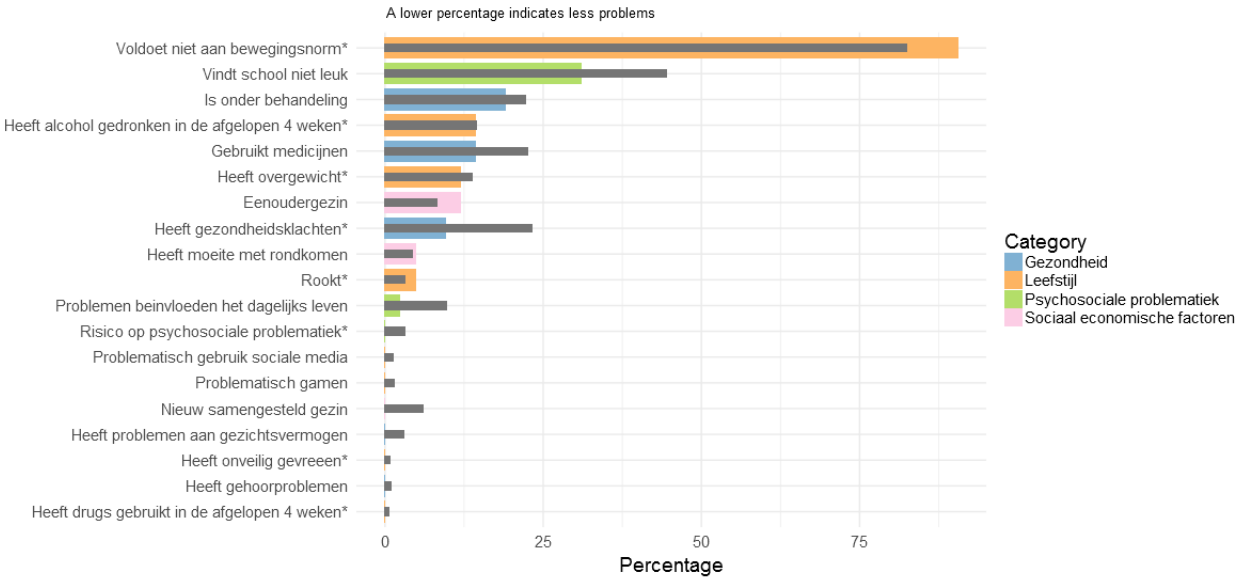

Figure 4.2.3. Variables sorted on descending percentages for the municipality Baarle-Nassau

In Figure 4.2.4, the data is presented in a different way. Here the in percentage differences of the values of municipalities with respect to the average is shown. Since we saw in Figure 4.2.2 that Baarle-Nassau performs very well, we see that it has a high deviation of the average from several variables. For example, we see that Baarle-Nassau performs very well on psychosocial problems, compared to the average of both regions.

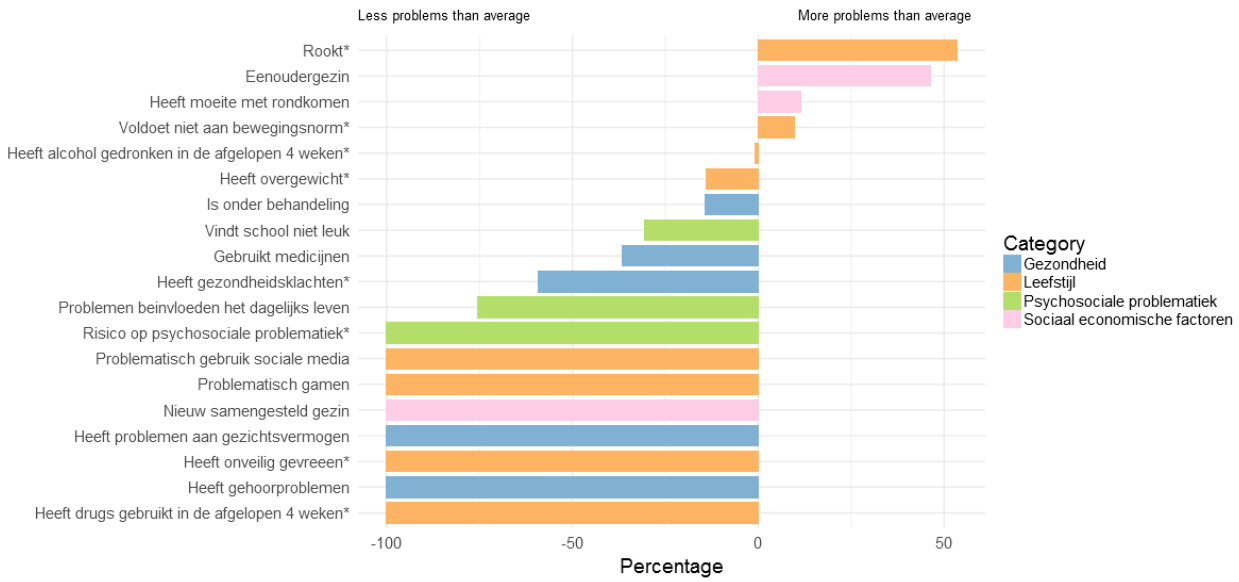

Figure 4.2.4. Variables sorted on deviation from the average for the municipality Baarle-Nassau

4.3 Trend Analysis

In this project, we focused mainly on the Hart voor Brabant region for trend analysis. With the agreement of GGD, we focused on selected variables that are indicators of the lifestyle of these students. Thus using GGD's domain expertise, we identified ten variables that are in the "lifestyle" category. The trend analysis is for all the municipalities in the region compared to the mean of the region itself, and is for the last three school years (2013-2014, 2014-2015, and 2015-2016). The names of these selected variables are given in Table 4 in Dutch and English.

We discuss first the overall trend of the Hart voor Brabant region, next as an example we take a municipality and compare it with the region. The trend analysis is done on percentages, and these percentages are calculated for problematic conditions. These problematic conditions are defined by GGD. An example of definition of a problematic condition is given in Chapter 3 (section 3.4) for the variable "gewicht5cat" that describes the overweight condition.

Based on this problematic condition definition, in the following we present the results for Hart voor Brabant for all the variables. For each variable, the weighted averages (means) are calculated using the sample size of each municipality as weights. Figure 4.3.1 shows the trends of all the variables in bar plots. The mean sample size over the three years is almost the same in the region. Over the last three years, the number of students who use drugs (Heeft drugs gebruikt in de afgelopen 4 weken), who smoke (Rookt), and who are being at risk of psychosocial problems (Risico op psychosociale problematiek) has a decreasing trend. This is a positive trend for the region as a whole. On the contrary, there is an increasing trend related to overweight (Heeft overgewicht) over the last three years. With respect to moving (exercising), which is given by the variable "Voldoet niet aan bewegingsnorm" most students do not meet the standard requirement and the percentages are high (> 75%) in all the years. In the years 2013-2014 and 2014-2015, the calculated averages are almost the same (~75%), but in the last year (2015-2016) it even increased to 81.5%. This can be related to an increase in percentages with respect to overweight, but we believe more research also should be done. For the other variables (alcohol and gambling), there is a decreasing and then increasing trend. In general, the largest deviation from requirement set is lack of physical exercise, as shown in figure 4.3.2.

Table 4. Variables used for trend analysis that are related to the lifestyle of students. This table gives the variable name in the dataset, the Dutch description for the variable and an English translation used in the project.

Variable name Dataset	**Dutch Translated Variable**	**English Translation variable**
alco4kwn	Heeft alcohol gedronken in de afgelopen 4 weken	Had drunk alcohol in the past 4 weeks,
drugs4wk	Heeft drugs gebruikt in de afgelopen 4 weken	Used drugs used in the past 4 weeks
beweeg4cat	Voldoet niet aan bewegingsnorm	Does not meet standard movement
gewicht5cat	Heeft overgewicht	Being overweight
gok4wk	Heeft gegokt in de afgelopen 4 weken	Gambled in the past 4 weeks
Gzhklacht	Heeft gezondheidsklachten	Has health problem
Nuroker	Rookt	smoker
sSDQcat	Risico op psychosociale problematiek	Risk of psychosocial problems
Schuld	Heeft schulden die niet binnen een maand kunnen worden afgelost	Has debts that cannot be repaid within one month
Vrijen	Heeft onveilig gevreeën	Had unprotected sex

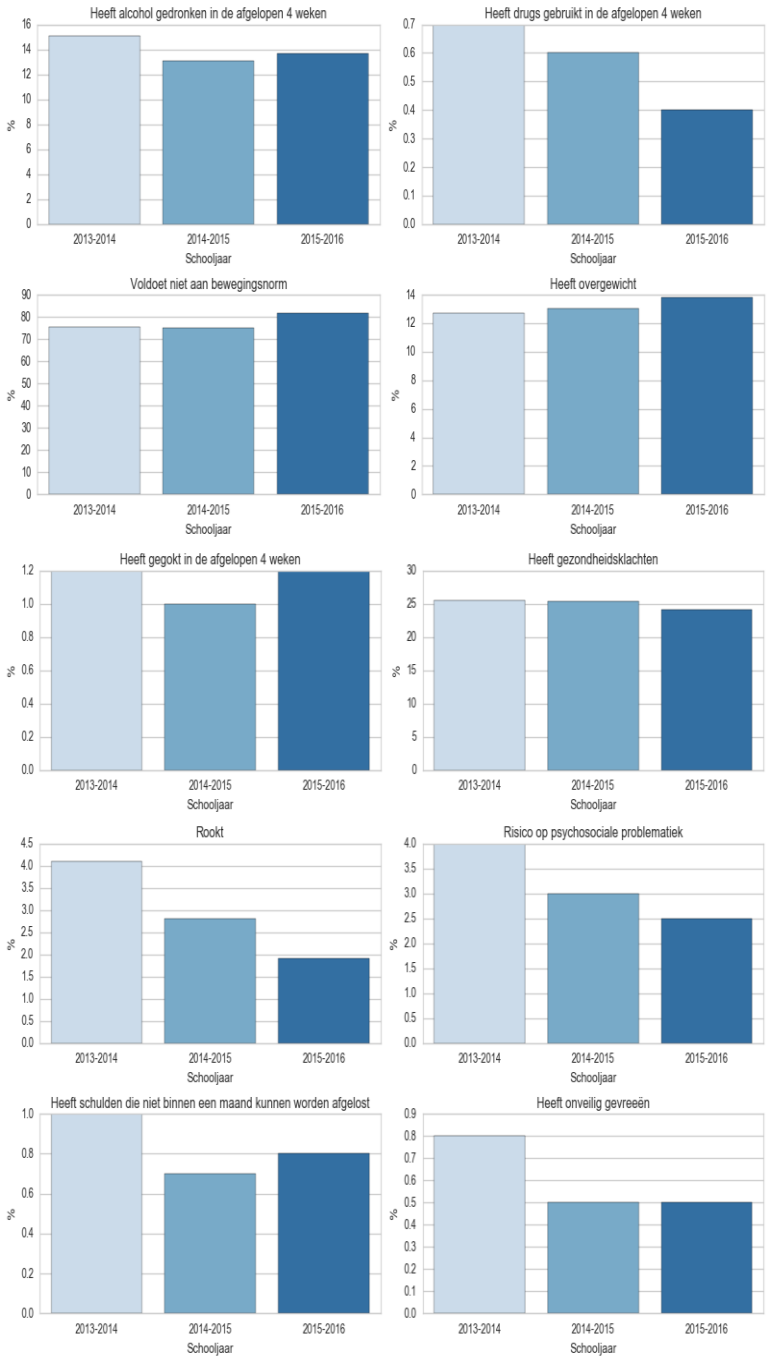

Figure 4.3.1. Trends of variables in the lifestyle category for the school years (2013-2014, 2014-2015, and 2015-2016).

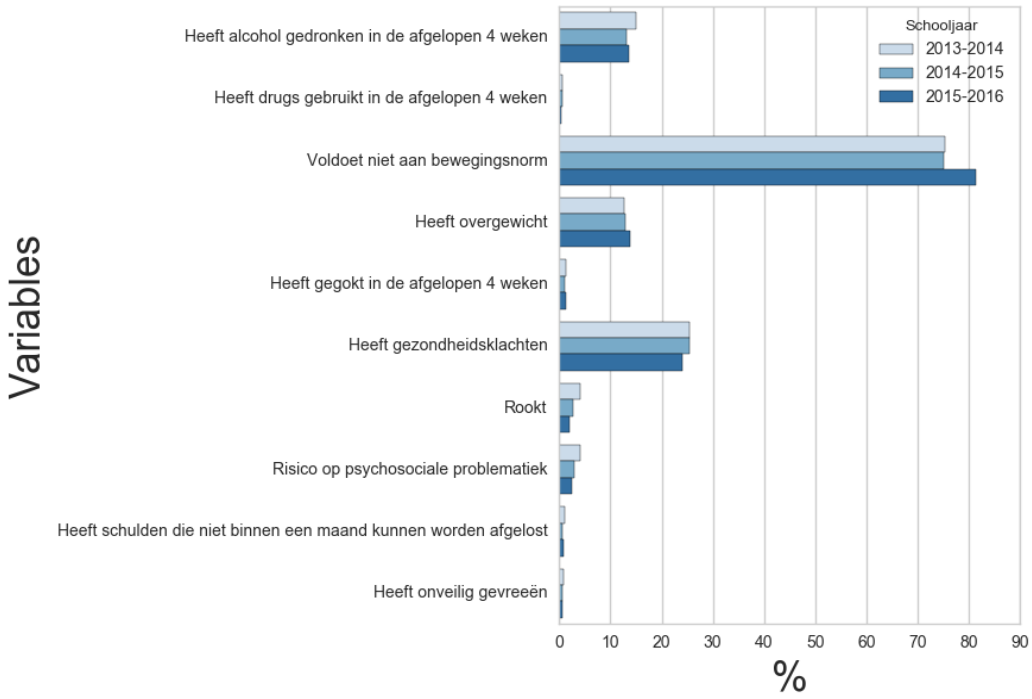

Figure 4.3.2. The percentages the life style variables for the school years (2013-2014, 2014-2015, and 2015-2016).

A similar analysis can be done for each individual municipality. To this end, the interactive tool we developed can help to identify the trends of these lifestyle variables for each municipality compared to the mean of the overall HvB region. Figure 4.3.3 shows the trend of the variable "Rookt": the percentage of students who are smokers in the municipality of 's-Hertogenbosch (in red) and the overall region of HvB (in green). From this we can easily see that the percentage of smokers has a decreasing trend in both the region (HvB) and in the municipality 's-Hertogenbosch. Although the values are almost the same over the years, the percentage of smokers in

's-Hertogenbosch are less in both 2013-2014 and 2015-2016 school year, and slightly higher in 2014-205. Other variables for each municipality compared to the mean of the region can be accessed from the visualization tool.

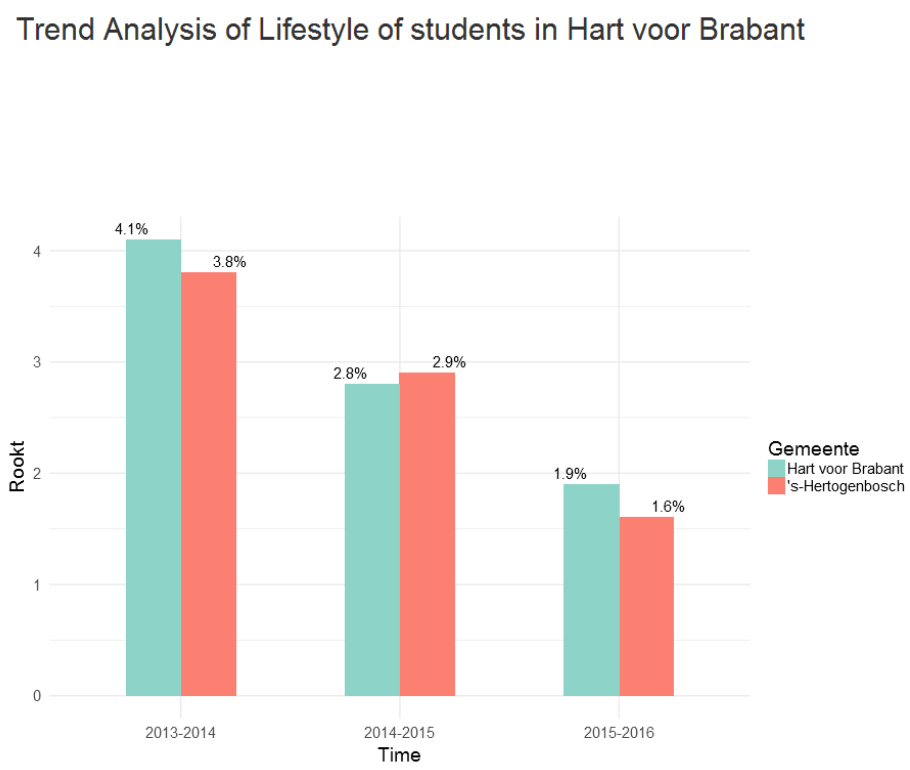

Figure 4.3.3. Bar plots for the trend of smokers in the municipality of 's-Hertogenbosch (in red) compared to the whole Hart voor Brabant region (in green) in percentages for the school year 2013-2014, 2014-2015, and 2015-2016.

5 Conclusions

- The pattern analysis option embedded in the interactive tool enables the benchmarking of municipality in respect to one or two variables. It also gives insights in the associations between variables.
- Using the interactive tool we developed, the profile of each municipality with respect to the variables of interest can be accessed easily.
- Although a better trend analysis needs more data, we conclude that the analysis we made gives some insight with respect to the lifestyle of the students.
- From the trend analysis we conclude that, a high percentage (> 75percentage) of the focus group, does not meet the standard of exercising according to the definition of GGD; this can be related to the increase in the overweight in the Hart voor Brabant.

6 Recommendations

- We recommend that GGDs use open source codes such as R for developing interactive visualization of data that can facilitate data exploration and decision-making.
- We have the following suggestions to extend the tool
    - Scaling the trend analysis for West-Brabant
    - Make an analysis on district level
    - Perform a benchmark between school types (VMBO/HAVO/VWO)
    - Plot problematic results on a map to search for patterns in areas.
- We recommend using the tool for new datasets, data cleaning processes should be generalized.
- We recommend collecting data over many years. It will improve the trend analysis over any variable
- We recommend that GGDs use the results, and especially identify outliers to inform policy makers.

Appendix A

List of variables used in the exploration of data, their description and the set threshold for problematic areas.

#	Variables	Description	Problematic threshold
1	gewicht5cat	Heeft overgewicht	4, 5
2	Visus	Heeft problemen aan gezichtsvermogen	2
3	Gehoor	Heeft gehoorproblemen	2
4	sSDQcat	Risico op psychosociale problematiek	3
5	simpactcat	Problemen beïnvloeden het dagelijks leven	3
6	gzhklacht	Heeft gezondheidsklachten	1 (yes)
7	medicijn	Gebruikt medicijnen	1 (yes)
8	behandnu	Is onder behandeling	1 (yes)
9	nuroker	Rookt	1 (yes)
10	gameind	Problematisch gamen	1 (yes)
11	Socind	Problematisch gebruik sociale media	1 (yes)
12	beweeg4cat	Voldoet niet aan bewegingsnorm	1,2,3 (Inactive)
13	eenouder	Eenoudergezin	1 (yes)
14	nieuwgezin	Nieuw samengesteld gezin	1 (yes)
15	meningschool	Vindt school niet leuk	3 (negative)
16	Vrijen	Heeft onveilig gevreeën.	1 (yes)
17	alco4wknnew	Heeft alcohol gedronken in de afgelopen 4 weken	1,2,3,4
18	nurondkomen	Heeft moeite met rondkomen	2, 3
19	drugs4wk	Heeft drugs gebruikt in de afgelopen 4 weken	1 (yes)