Professional Doctorate in Engineering

Data Science

Visualization of the West-Brabant and Hart for Brabant GGD Adults Monitor Data

Blagorodna Ilievska

Eleftherios Koulierakis

Berisha Mekayhu

Diego Perez

February 17, 2017

PDEng Program Data Science

Jheronimus Academy of Data Science

Eindhoven University of Technology

Convent Mariënburg

Sint Janssingel 92

5211DA 's-Hertogenbosch

Executive summary

Approximately every four years, each GGD sends questionnaires to the residents of its region to monitor the well-being and health of the people in the municipalities associated to that GGD. The GGDs split the residents into two groups according to their age:

1. Adults (people who are 19 years old or older and younger than 65)
2. Elders (people who are 65 years old or older).

A group-specific questionnaire is given to each age group. Some questions may be the same for both groups. The focus of the questions for Adults is on themselves and their surroundings, while for Elders, it is on their health and their ability to support themselves. The responses of all the questionnaires are gathered in a database, which is called the Monitor data.

This data is used by the policy makers and policy advisors to generate and evaluate measures at municipality level. They often are interested in data explorations, asking questions, such as 'What is the most striking health indicator in a certain municipality?'. To help such explorations and answer these questions, the GGDs want to improve the ways in which they visualize the Monitor data, in order to assist policy makers in the best way possible. The goal of this project is to provide an interactive visualization approach that can support the decision making process for policy makers. In order to achieve this goal, we developed a dashboard that includes the following visualization tools:

- The first tool applies the concept of an I^f-F^I table, which enables users to deal with high-dimensional data using the well-known table metaphor. In the first table, municipalities are displayed as rows and the features are displayed as columns, where cells contain values of features per municipality. In the second table, features are displayed as rows and the municipalities are displayed as columns, where cells contain municipality values per feature.
- The second tool is a map, on which the values of the selected feature in the I^f-F^I table are displayed at municipality level.
- The third tool is a scatter plot, which shows interactions between variables. In our case, the scatter plot is shown for the two selected features in the I^f-F^I table.

The three tools facilitate quick look-up of exact values. By clicking on a municipality, all attribute values for that municipality are immediately displayed. By clicking on an attribute, the attribute values for every municipality are displayed. These values are also mapped on municipality level, and if two attributes are selected a scatter plot is shown.

We presented a new method for the exploration of high-dimensional multivariate data using tables, extended with a map and a scatter plot.
We recommend installing and using the tool on a trial basis. If the trial tool proves successful, we recommend further development which will include the Monitor 2016 data, data exploration on district level and data exploration on regional level.

Table of Contents

1. Management introduction 7
   1.1 Problem background 7
   1.2 Goal specification and added value 8
   1.3 Strategy 8
   1.4 Results 9
   1.5 Conclusions 10
   1.6 Recommendations 10
2. Stakeholder analysis 11
2.1 Primary collaborators 11
   2.1.1 GGD West-Brabant and GGD Hart voor Brabant 11
   2.1.2 JADS (TU/e) 11
   2.2 Secondary collaborators 11
   2.2.1 Executives 11
   2.2.2 Technical supervisor 12
3. Data value chain 13
   3.1. Data source and value 13
   3.2 Dataset information 13
   3.3 Tasks to be performed 15
   3.4 Data Science strategies 15
   3.5. Literature review 16
   3.6. List of variables, data, and abbreviation 16
4. Dataset characteristics, structure and analysis 19
   4.1. Entity relation diagram and attributes 19
   4.2 Type of data and scale 19
   4.3. Data manipulation 19
5. Results 21
   5.1 I^f-F^l Table Functionality 21
   5.2 Scatter Plot Functionality 23

5.3 Map Functionality 25
5.4 Use Case 25
6. Conclusions 30
7. Recommendations 31
Future Work 31
References 32
Appendix 33
Used Software 33

List of Tables

Table 3.6.1: Sample List of Variables 17
Table 3.6.2: Structure of provided dataset 18
Table 3.6.3: Structure of the supplementary answers dataset 18

List of Figures

Figure 5.1: Quantitative representation of the dataset in the I^f-F^I Table 22
Figure 5.2: Representation of the I^f-F^I Table, map and scatter plot, when two municipalities and two features are selected..... 23
Figure 5.3: Sample scatter plot for Autochtoon vs Vrouw..... 24
Figure 5.4: Representation of the Map Functionality..... 25
Figure 5.5: Initial representation of the Visualization..... 26
Figure 5.6: Selecting the city of 's-Hertogenbosch..... 26
Figure 5.7: Selecting the city of 's-Hertogenbosch..... 27
Figure 5.8: Comparing the Municipalities of 's-Hertogenbosch and Tilburg. 27
Figure 5.9: Comparing the alcohol consumption percentages in different municipalities..... 28
Figure 5.10: Comparing the alcohol consumption and obesity rates in different municipalities.28
Figure 11: Obesity rates against alcohol consumption..... 29

1. Management introduction

1.1 Problem background

The GGD GHOR is the Association of GGDs (Community Health Services) and GHOR (Regional Medical Emergency Preparedness and Planning) offices in the Netherlands. This governmental institution serves as the focal point for the public health efforts in The Netherlands. The working area is very diverse, but in its core its duty is to look after the interests of the GGDs and the GHOR offices, by guarding, improving, and promoting public health.

The Netherlands has 25 local community health services (GGDs) that provide coverage across the country. The main responsibility for local community health services is preventive health care. They monitor health risks and try to improve people's health by taking the necessary steps to mitigate health risks. In order to react effectively to health issues, the GGDs have to monitor and distinguish potential factors that may affect the citizens' well-being.

One of the ways in which the GGDs are able to understand the needs of the communities is by the Gezondheidsmonitor (Health Monitor). This survey collects systematic and uniform health information for all municipalities, in such a way that comparison with national and regional figures is possible. The questionnaires intend to map the health of the citizens on certain domain specific features by collecting information on health, lifestyle and environment. The survey is conducted every four years for a sample of about 25,000 citizens in both regions.

The Public Health Act places responsibility for health policy at the local councils. Municipalities support their decisions with data on the health of their inhabitants collected by the Health Monitor survey. To properly use this data, there is a need for a tool that can efficiently expose and summarize the relevant factors associated to each municipality, district or region. Currently, the GGDs can make the Monitor data available to policy makers by means of

a Tabellenboek (table book) and factsheets. These tables aggregate the totality of the monitor survey results, but are hard to read and interpret, because of the nature of the data. Better mechanisms of communicating and exploring the monitor data could result in better policies associated to the municipality's health issues.

1.2 Goal specification and added value

With the current solution, the data is readily available to decision makers in its totality. Nevertheless, scanning through tables is inefficient because of the relatively large number of variables and the fact that all data is presented. It is difficult to draw conclusions from the tables because extracting knowledge out of them requires thorough examination. The goal of the assignment is to propose a software tool that makes this process more efficient. Additionally, the current way in which the information is presented, serves to confirm or deny pre-conceived hypotheses. In this context, the purpose of this study is to create a tool that allows the information to speak for itself without the need of much additional input. In the desired situation decision makers ae enabled to explore the data without having a particular hypothesis in mind. In the current situation, benchmarking between indicators or geographic locations is a laborious process. For this reason, the proposed tool should enable decision makers to see correlations between indicators and recognize patterns. Currently, the data is visualized on the GGD's website. In this visualization, after the user has chosen a feature of the survey, is presented in a table or a map. Even though this functionality allows for exploration of the data, the user has to have a feature (or features) that he deems relevant or important in mind. It is desired that the proposed visualization presents the data in an attractive and interactive format.

Finally, the solution should be scalable for different geographical divisions (such as district, municipalities or regions) and new questionnaires for upcoming surveys.

1.3 Strategy

Since this study focuses on local policy making, the starting point for the requirement collection of this project was to create a list of questions that local policy makers or advisors might want

the Monitor data to answer. An extensive list is presented in Section 3.2. With the use cases in mind, brainstorming sessions were carried out to propose potential visualization tools that could answer the questions posed in the collected use cases. After selecting the correct set of tools that serves the purposes, the implementation tasks were divided among the team. The next step was integrating the proposed tools in such a way that there is synergy among them. Meanwhile, the data was cleaned and aggregated to be compatible with the selected visualization tools. Lastly, with a visual design in mind, a dashboard was created. This last step includes the selection of the appropriate design elements that make the tool user friendly.

1.4 Results

Having the use cases in mind, we produced a list of functional requirements that the software tool should have. After proposing different methods to tackle the needs of the users, three techniques were selected: I^f-F^I tables, map tool, and scatter plot.
The spine of our dashboard is the I^f-F^I table. This tool allows for exploration of municipalities and features for high dimensional data. The visualization consists of two juxtaposed tables: an I^f-Table, showing all items for a selection of features; and an F^I-Table, showing all features for a selection of items. Thus we enable the user to limit the number of visible items and features to those needed for the exploration.

Extending this functionality, the map and scatterplot present ways in which geographical and feature comparisons are possible. The capacity to compare municipalities given a set of features is included in the visualization dashboard via a scatter plot. For the scatterplot the data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis. This type of exploration is useful to investigate trends, potential correlations and patterns in the interaction of features and to detect outliers. The map is a visual representation of an entire area or part of an area, represented on a flat surface. This functionality is a straightforward method that supports geographical comparisons.

The three tools are then integrated into one dashboard. By the aggregation and synergy of the three tools we are able to address the research questions presented before.

1.5 Conclusions

The proposed I^f-F^I tables allow for the exploration of high dimensional data and present an intuitive way to explore tabular data. Extended with map and scatter plot functionalities, it helps to benchmark municipalities and features. The produced tool is interactive and useful to discover patterns within municipalities and features. Finally, the tool helps to explore the relations and interaction between pairs of features.

1.6 Recommendations

We recommend that the GGDs and their end-users integrate the visualization tool into their IC technology to explore the monitor data. We suggest that users get acquainted with the software and use the tool on trial basis. For further work, we suggest that the 2016 monitor data is explored with help of the visualization tool. Additionally, we suggest that two extensions are added: aggregation on district and regional level and implementation of municipality and feature relevance and similarity.

2. Stakeholder analysis

2.1 Primary collaborators

2.1.1 GGD West-Brabant and GGD Hart voor Brabant

Twenty five GGDs form the health organization for preventive healthcare in The Netherlands. The GGDs have several focus groups, among which children, travelers and senior citizens. Their main tasks are monitoring, advising and informing clients about medical care. The almost 400 municipalities in the Netherlands are taken care of by the 25 GGD health service communities. The tasks of the community health services are not always identical, since every municipality gives its own assignments to the GGD. For this project, two GGDs are involved: GGD West-Brabant and GGD Hart voor Brabant.

2.1.2 JADS (TU/e)

JADS is an abbreviation of Jheronimus Academy of Data Science. JADS is a joint initiative of Eindhoven University of Technology, Tilburg University, the Municipality of 's-Hertogenbosch and the province of Noord-Brabant. The aim of JADS is to train the next generation of data scientists and entrepreneurs and become a center of knowledge and activity, where future-proof, real-world solutions are developed. An important mission of JADS is to understand and to advance the value of data in solving complex societal and business challenges.

2.2 Secondary collaborators

2.2.1 Executives

The representatives of GGD and problem owners are Wieteke de Vries, Linda van Tilburg, and Anne Groet. They work as researchers at the research department of both GGDs, West-Brabant

and Hart voor Brabant. Their goal is to provide knowledge about the health of all inhabitants in their regions. To do so, they use both quantitative and qualitative methods. Their quadrennial questionnaires provide the largest part of the information. The data they collect is used by policy makers and policy advisors to base their municipal policies on. They often ask the GGDs questions about the data such as 'What is the most striking health indicator in a certain neighborhood?'. To help them answer these questions, they have the desire to improve the visualization methods of their data in order to assist policy makers in the best possible way.

2.2.2 Technical supervisor

The technical supervisor of the project is Jarke J. (Jack) van Wijk. Jack is a professor of visualization at the Department of Mathematics and Computer Science of Eindhoven University of Technology (TU/e). He is also the scientific director of the Professional Doctorate in Engineering in Data Science program.

3. Data value chain

3.1. Data source and value

The datasets are provided by the public health institution (GGD), which provides diverse health services in the Netherlands. For this project, we used the 2012 monitor data from the GGDs, Hart voor Brabant (HvB) and West-Brabant (WB). The monitor data is a combination of the 2012 survey data and the data from Statistics Netherlands (Centraal Bureau voor de Statistiek, CBS). There is no overlap between the respondents who filled out the GGD questionnaires and those who filled out the CBS questionnaires. To segment the population according to age: Adults are residents who are older than 18 and younger than 65, and Elders are residents who are older than 64.

The survey outcomes and CBS data are used to monitor, protect, and promote the health of the citizens of the two regions, with special attention to risk groups. Presenting and analyzing this data in a form that supports decision making for policy advisors and researchers is one of the core tasks of GGD. The data collected from the residents of the municipalities is reported to the policy advisors, local policy makers, and concerned entities in the municipalities.

Currently the way to access the monitor information is by the Tabellenboek (table book) and factsheets. These tables aggregate the totality of the monitor survey results, but are hard to read and understand. Understanding the monitor data and its patterns facilitates the decision making in health and social issues. Better mechanisms of communicating the monitor data could result in more effective policies associated to the municipality's health issues.

3.2 Dataset information

Approximately every four years, each GGD sends questionnaires to the residents of its region to monitor the well-being and health of the people in the municipalities associated with that GGD.

The dataset contains the data that was collected with a survey from representative samples of the residents in the regions. The Monitor data is at an individual level and is stored in both SPSS and Excel files, and spreadsheets where each row corresponds to one completed questionnaire. The variables in each dataset correspond to the items of the questionnaire and to their outcome after processing by the GGD. With few exceptions, the variables in the Monitor data are categorical data. Processing the data about the health of the residents in all age groups and delivering the results to concerned bodies is one of the tasks of a GGD.

Communicating the results and conveying the information from the monitor data requires improvements of the current tools, such as interactivity and easy visualization. The business case is to support the decision making, in such a way that end users can effectively use the monitor data.

The project aims to produce a visualization solution that addresses the difficulty of understanding the volume of the monitor data. The visualization solution aims to improve the access of information by end-users. The solution should support the task of comparison of health indicators among municipalities, and explore hypotheses on the data.

The rapid change in the role of local government in social services, the role of citizens in helping each other, and the changing needs in society urge modernization of the GGD's services. The general objective of GGD in this project is coming up with a data driven solution that improves the current services related to health monitor data. Specifically, the two GGDs wish to answer the following questions:

1. Can we visualize the monitor data in a more interactive way?
2. Can we develop a visualization that can be easily understood by our end-users?
3. Can we develop a scalable visualization solution suited for new datasets?
4. Can we easily compare municipalities on specific health indicator(s)?
5. Can we easily compare health indicators in certain municipalities?
6. Can we see useful correlations and patterns in the monitor data?

3.3 Tasks to be performed

We define the data visualization tasks that can answer GGDs' questions. In the first step, we started by listing possible questions that end users may ask. The main questions are:

1. Which health indicator is important for a certain municipality?
2. Which municipality has a high/low value for given indicators?
3. For which areas do the indicators behave similarly?
4. For which indicators do municipalities behave similarly?
5. Are there significant correlations/patterns among indicators within the region?

Considering the end-users, the second task was to answer these interesting questions step by step. Explicitly, the tasks were cleaning and aggregation of data, implementation of the I^f-F^I tables, and then extend the table functionality with a scatterplot and a map visualization. We designed and developed a visualization tool that interactively allows for comparisons and observation of relationships in the monitor data.

3.4 Data Science strategies

The strategies to accomplish the above tasks take into account the current situation of GGD in using monitor data. We designed a data visualization dashboard that displays all features and municipalities. The nested model of visualization design (Tamara Munzner, 2009) was used to design the visualization tool . This model is defined in four nested layers. The first outer layer is the domain problem characterization, the second layer is data/operation abstraction design, the third layer is encoding/interaction technique design, and the fourth inner layer is algorithm design.

The visualization tool is comprised of I^f-F^I tables in combination with a map and a scatterplot. The dashboard displays these data visualization tools in one screen and provides ample interactions.

Preparation of the data was the first activity in the implementation phase of the project. After this we implemented the tables, map and scatterplot independently. Finally, we combined the three results into one visualization tool, using the R Shiny package. The expected result is the dashboard tool, which displays the I^f-F^I tables, that help compare municipalities for specific health indicator(s) and vice versa; a map that displays the selected indicators geographically; and a scatter plot that compares two indicators, where direction, shape, and strength of correlations between indicators are shown.

3.5. Literature review

The exploration of high-dimensional data is challenging because humans have difficulty to understand more than three dimensions. Tables, scatter plot matrices (SPLOM), parallel coordinate plots (PCP), and pivot tables are the standard workhorses for the exploration of multivariate data. Recently, Van der Corput and Van Wijk (2016) introduced a new visualization concept called I^f-F^I tables, which stands for Features-Items and Items-Feature tables. Using this visualization concept, comparison and observation of relevance and similarities among items and features can be explored. According to van der Corput and Van Wijk, the integration of different visualization techniques such as scatterplot and data reduction techniques like PCA with an I^f-F^I table could improve the insight of visualization. The map is a useful tool used to make quantitative comparisons among geographic locations. The map is useful to monitor health related issues in different locations, and it helps policy makers to target on geographic proxies. Additionally, the scatter plot extends the previous functionalities to suggest various kinds of interactions between variables.

3.6. List of variables, data, and abbreviation

The datasets are collected using a set of questions under 12 different themes; these are: general theme, financial situation, perceived health and well-being, physical health, mental health, life style, annoyance by physical environment, residential area, social environment, social safety,

curative care, and services. The dataset was presented as a spreadsheet: each row corresponds to a respondent and each question corresponds to a column. Each entry of the spreadsheet indicates the value/code answer of the respondents. A sample list of variables is presented in Table 3.6.1

Table 3.6.1: Sample List of Variables

Table 3.6.1: Sample List of Variables

Theme	**SPSS name**	**Label**	**Code Values**	**Value Meaning**	**Questionnaire Name**
General	geslacht	sex	1	Man	AGGSB201
			2	Vrouw	
			9	NA	
Physical Health	obesitas	obesity	0	nee	AGGWB201, AGLNB201
			1	ja	
			9	NA	
Perceived health and wellbeing	KLGGA208	Overall health in 2 categories	0	Goes well, bad to very bad	KLGGB201
			1	Very good to good	
…….	….	…….	…	…	…

As explained above, the provided lists in the dataset have the structure shown in Table 3.6.2. The questionnaire data has one row per respondent and one column per question. Each of the cells encodes the answer of a particular person to the corresponding question.

Table 3.6.2: Structure of provided dataset

Respondent	Gemcode	Survey number	Question 1	Question 2	...
1	284	2	0	1	
2	846	1	1	2	
...					

The answer encoding was provided in a separate file. The structure of the supplementary dataset that encodes the answers has the structure shown in Table 3.6.3. This file contains the type (binary, categorical, or continuous), possible values for each answer, the meaning of each value, the short name of the answer, the long name of the answer, and a column that indicates if the variable should be included in the visualization. The short and long names contain the text that is going to be displayed in the visualization.

Table 3.6.3: Structure of the supplementary answers dataset

Question: SPSS name	Type	label	Value	Value Meaning	Short Name	Long Name	Used
geslacht	binary	sex	1	Man	Man	Geslacht: % mannen	x
geslacht	binary	sex	2	Vrouw	Vrouw	Geslacht: % vrouwen	x
geslacht	binary	sex	9	NA			
leeftijdcat4	categorical	age in 4 categories	1	19-24	19-24 jaar	Leeftijd in 4 categorieën: % 19-24-jarigen	x
leeftijdcat4	categorical	age in 4 categories	2	25-39	25-39 jaar	Leeftijd in 4 categorieën: % 25-39-jarigen	x

leeftijdcat4	categorical	age in 4 categories	3	40-54	40-54 jaar	Leeftijd in 4 categorieën: % 40-54-jarigen	x
...							

4. Dataset characteristics, structure and analysis

4.1. Entity relation diagram and attributes

The data was collected from sample respondents from each municipality in the HvB and WB regions. The data about health and social care for the two survey groups was gathered via questionnaires; a sample of adults and elders. According to the GGDs, the sample respondents were selected on proportion of population in age groups in districts. A group-specific questionnaire is given to each age group. Some questions may be the same for both groups. The focus of the questions for Adults is on themselves and their surroundings, while for elders, it is on their health and their ability to support themselves.

4.2 Type of data and scale

The monitor dataset is a multivariate cross-sectional dataset in tabular format. The dataset is in two Excel tables, one is the data for Adults and the other is the data for Elders. The adults' dataset contains 1772 columns/variables and 24,642 rows/respondents. The elders dataset has 1354 columns/variables and 24,601 rows/respondents. For this project we took samples that correspond to the GGD survey. In this case, the columns/variables in the table represent items in the questionnaire. The number of items represents people who filled the questionnaire.

4.3. Data manipulation

The dataset presented in Table 3.6.2 was cleaned and aggregated so that the resulting dataset has the structure shown in Table 4.3. In this aggregated dataset, there is one row per municipality and one column per possible answer to all of the questions. Each of the cells contains the percentage of respondents of the municipality that responded with the corresponding answer.

The decision to aggregate the data at municipality level comes naturally from the target users. For each of the datasets, each column/respondent has a municipality code associated. Different geographical aggregations can be made. Moreover, one of the main use cases was centered on the possibility of empowering policy makers at a municipal level.

The team explored the possibility of aggregating at a district and regional level. These aggregation levels presented problems since aggregating at district level requires techniques that handle the fact that one district can be part of two or more municipalities. For this case it would be necessary to distribute the district data based on the percentage of the district that is on each municipality. We understand that aggregating at a district level is useful, but the proper simulation of this data falls out of the scope of this exploratory visualization project. Aggregation at a regional level is straightforward and could be useful to compare municipalities with their corresponding region, but due to time constraints, this aggregation was left as future work.

Table 4.3: Structure of the clean and aggregated dataset

Municipality	**Question 1, Answer 1.**	**Question 1, Answer 2.**	**Question 1, Answer 3.**	**...**
1	50%	20%	30%	
2	40%	45%	15%	
..				

The values of each possible answer are counted and aggregated to percentage at municipality level. Once the data has been aggregated in this way, the dataset is exported and ready to be used by each of the functionalities in the dashboard. The structure of the prepared dataset is important to understand, in case the dashboard should be reused with a different dataset. The shapefile (geographic polygon data) of 46 municipalities is used to map the aggregated health indicators.

5. Results

The problem of visualizing a multivariate dataset that contains information for different regions has been tackled by implementing the I^f-F^I tables. The I^f-F^I table is a simple, yet extremely efficient and straightforward visualization tool that allows users to focus on the attributes and the regions they are interested to study. We introduced additional features to the I^f-F^I table: The map and the scatter plot. Both features extend the capabilities of the I^f-F^I table. The map allows the exploration of geographic effects and the scatter plot uncovers the correlation between different features. Our visualization tool can serve several purposes in research projects, such as the study of alcohol consumption in various regions, the comparison of obesity rates, nutritional habits, and employment rates.

5.1 I^f-F^I Table Functionality

In order to visualize the Monitor data we use the bottom-up approach, which means we consider all municipalities and all features at the same time. Understanding this high-dimensional data at a low aggregation level is challenging, because humans have difficulty to comprehend more than three dimensions. We present the use of a concept that enables users to explore such data and, specifically, to learn about important items and features that are unknown or overlooked, based on the items and features that are already known. The visualization consists of two juxtaposed tables: an I^f-Table, showing all items with a selection of features; and a F^I-Table, showing all features with a selection of items. This enables the user to limit the number of visible items and features to those needed for the exploration. The table is by far the most popular method to present multivariate data. In our case, the first table consists of all municipalities, displayed as rows, the features (or the attributes) are displayed as columns, and the cells contain values of attributes per municipality. Features are displayed as rows in the second table, where the columns represent municipalities and cells contain municipality values per attribute.

Municipalities

Show 25 entries Search:

Municipalities	Midden.opgeleid.2 %	Gescheiden..levend. %
Aalburg	37	5
Baarle-Nassau	35	5
Bergen op Zoom	34	12
Boekel	45	3
Boxmeer	41	6
Boxtel	36	7
Breda	32	9
Dongen	38	4
Etten-Leur	42	6
Geertruidenberg	41	4
Gilze en Rijen	35	7
Goirle	37	9
Grave	37	6

Features

Show 25 entries Search:

Features	Breda %	's-Hertogenbosch%
Man	43	46
Vrouw	57	54
X25.39.jaar	29	30
X40.54.jaar	38	36
X55.64.jaar	25	28
X19.24.jaar	9	6
Autochtoon	78	86
Westers.allochtoon	11	8
Niet.westers.allochtoon	11	6
Midden.opgeleid.2	32	28
Midden.opgeleid.1	25	23
Hoog.opgeleid	36	44
Laag.opgeleid	7	5

Figure 5.1: Quantitative representation of the dataset in the I^f-F^I Table

Sorting items by the values of a feature by clicking on a column facilitates the fast selection of interesting items, while vertical scrolling facilitates linear scanning. The number of features shown is limited to two, to prevent the need for horizontal scrolling, and features can be selected separately, using a search mechanism.

Every time when a row in the Municipalities table is selected, a column is added in the Features table, where in every cell municipality inline bar plot values for each feature are displayed. Every time when a row in the Features table is selected, a column is added in the Municipality table, where in every cell feature inline bar plot values for each municipality are displayed. The number of selected rows is limited to two. When two rows in the Features table are selected, the scatter plot is displayed for these two features.

Figure 5.2: Representation of the I^f-F^I Table, map and scatter plot, when two municipalities and two features are selected

5.2 Scatter Plot Functionality

Benchmarking is an important requirement expressed by the problem owner. The capacity to compare municipalities given a set of features is included in the visualization dashboard via the scatter plot functionality. This type of plot is capable of exposing relations between features and showing the behavior of each municipality in any pair of features. Figure 5.3 presents as an example a scatter plot for Autochtoon vs % Vrouw.

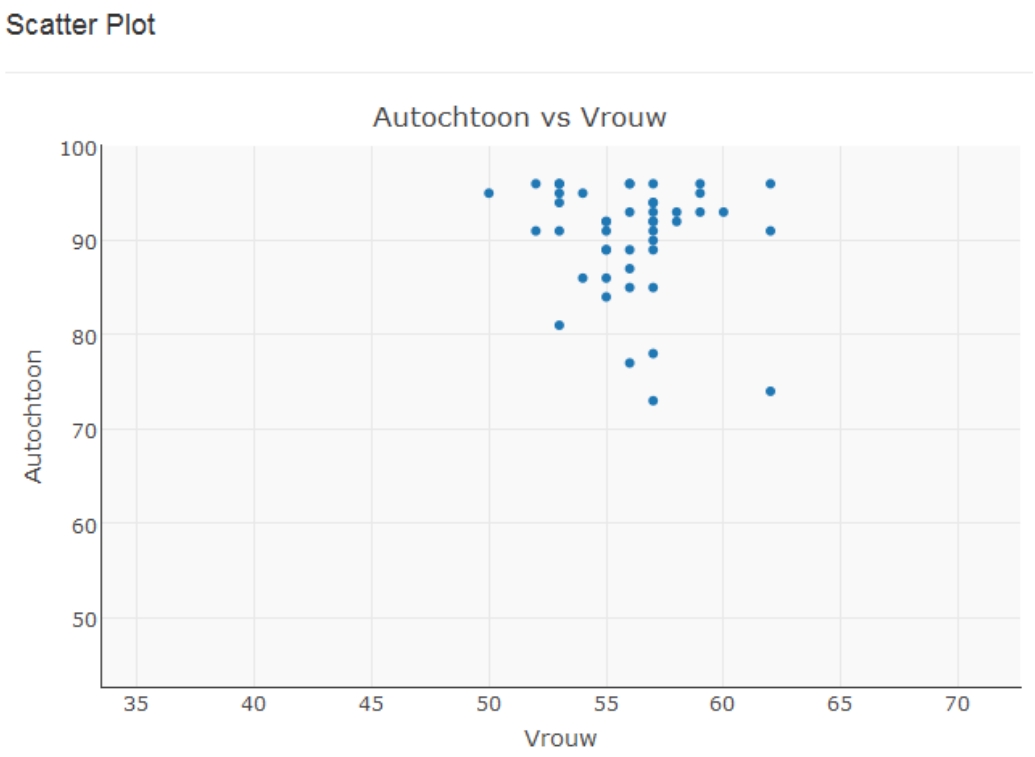

Figure 5.3: Sample scatter plot for Autochtoon vs Vrouw

The scatter plot is constructed by two parameters, namely the two selected features of the I^f-F^I Table. Once two features have been selected from the Features table, a scatter plot appears in the lower right corner of the dashboard. Once one feature is plotted against the other, the user can visually explore the interaction of the features and their relation. With this functionality, each point on the scatter plot represents one municipality. The user can hover over each individual point in the plot to expose a tooltip that indicates the municipality name for the user. This type of exploration is useful to investigate trends in the interaction of features and to detect outliers. Moreover, since the scatter plot was implemented with the R library Plotly, this plot is made interactive with the features of zooming, panning and hovering.

The purpose of the scatter plot is to tackle the limitations of tables for visually finding clusters and analyzing correlations between features. This plot provides a detailed view of the possible correlation of interesting features that are found in the tables.

5.3 Map Functionality

An important feature introduced in our visualization tool is the map functionality. In order to develop the map and to integrate it to the I^f-F^I tables we used R and Shiny. We used a scalar color range to display the percentage of values on each municipality. Every time the user clicks on a specific feature at the F^I table, all the municipalities are displayed with a color related to the percentage of people that answered positively to the question of the corresponding feature. Sometimes it is difficult to identify each municipality and for that reason we added a hover window that displays the name of the municipality and the percentage of the selected feature. The map is a useful tool that allows the user to make quantitative comparisons among municipalities and to explore possible similarities.

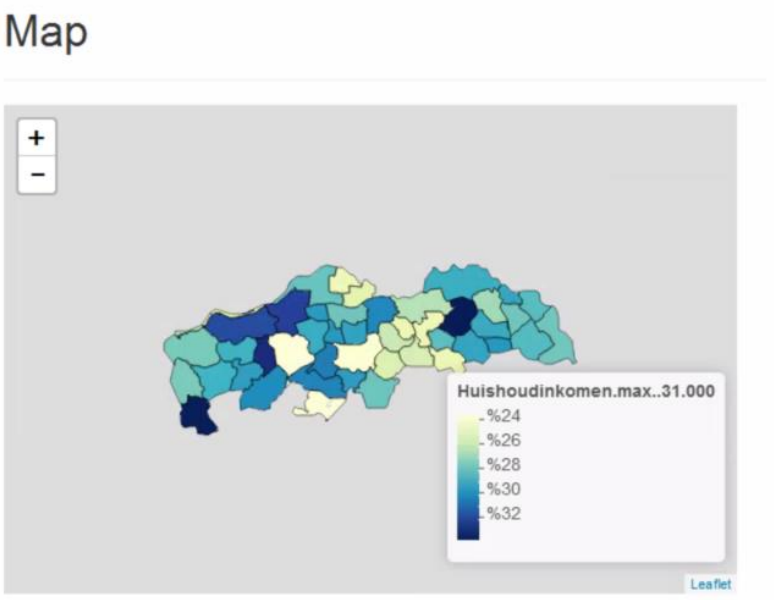

Figure 5.4: Representation of the Map Functionality

5.4 Use Case

Assume that we want to study the municipality of 's-Hertogenbosch, one of the most beautiful cities in the Netherlands. We launch the visualization tool and we obtain Figure 5.5.

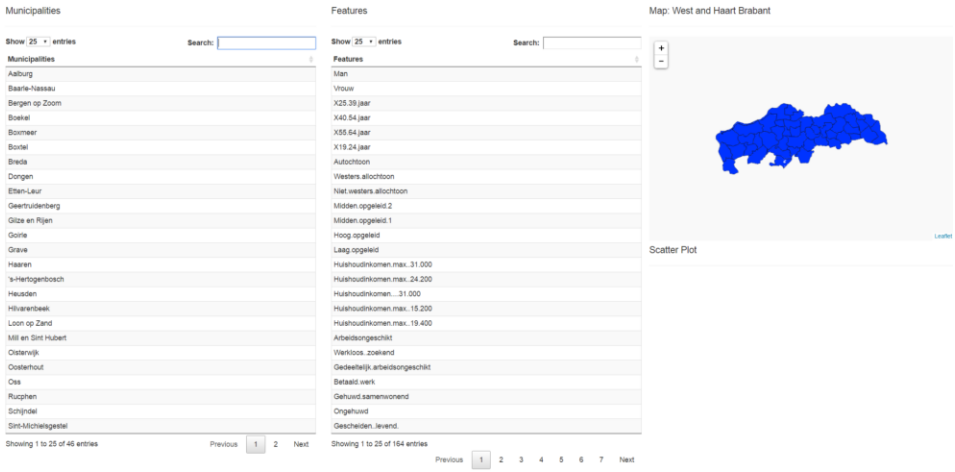

Figure 5.5: Initial representation of the Visualization

In order to find Den Bosch we use the search bar that is at the top left corner of the tool. We type '**s-H**' and we obtain 's-Hertogenbosch. We then select the municipality and we obtain the percentages for every attribute as shown in Figure 5.6.

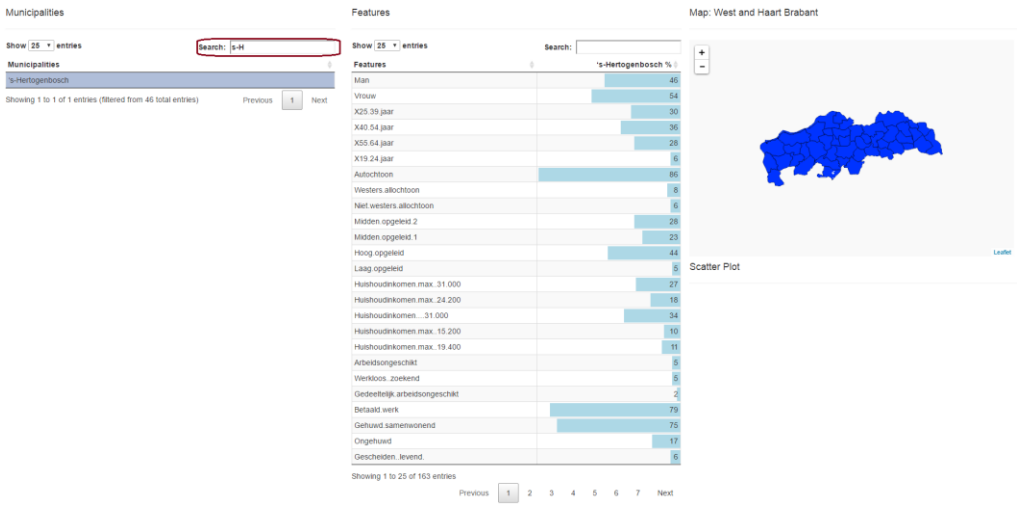

Figure 5.6: Selecting the city of 's-Hertogenbosch

Attributes can be sorted in ascending or descending order depending on their corresponding percentages by clicking at the sort option below the search bar of the attributes. In Figure 5.7 we sort the attributes in descending order.

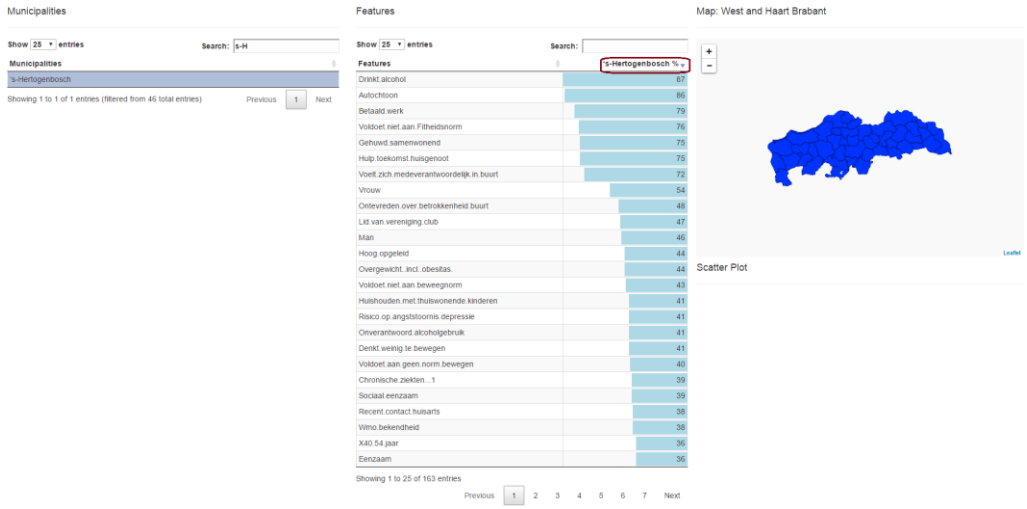

Figure 5.7: Selecting the city of 's-Hertogenbosch

Now we want to compare two different cities. Tilburg might not be as beautiful as Den Bosch, but it is also one of the main cities in the region of Brabant, so comparing these two municipalities will be interesting. In order to do so, we search for Tilburg at the top left search bar. By clicking '**til**' we obtain Tilburg and then we select it (Figure 5.8). Notice that the names of the two cities are displayed on top of the two columns on the F[l] table.

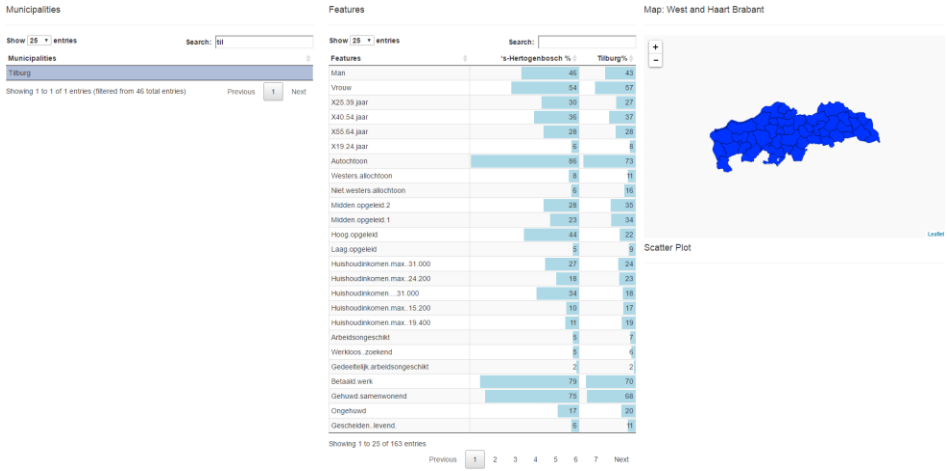

Figure 5.8: Comparing the Municipalities of 's-Hertogenbosch and Tilburg.

Right now we would like to see the alcohol consumption in every region. For that reason, we just search '**alco**' at the top right search bar and we select the feature with the name **Drinkt.alcohol**. We obtain the visualization of the percentages on the map. The color scale goes from light yellow (low percentages) to dark blue (high percentages). In order to find the municipality with the highest alcohol consumption, we sort the alcohol consumption column in ascending order by clicking to the sorting bar. We see that the municipality with the highest alcohol consumption is Vught. By observing the map we see that Tilburg has the lightest yellow

color, so the alcohol consumption at this municipality is the lowest (Figure 5.9).

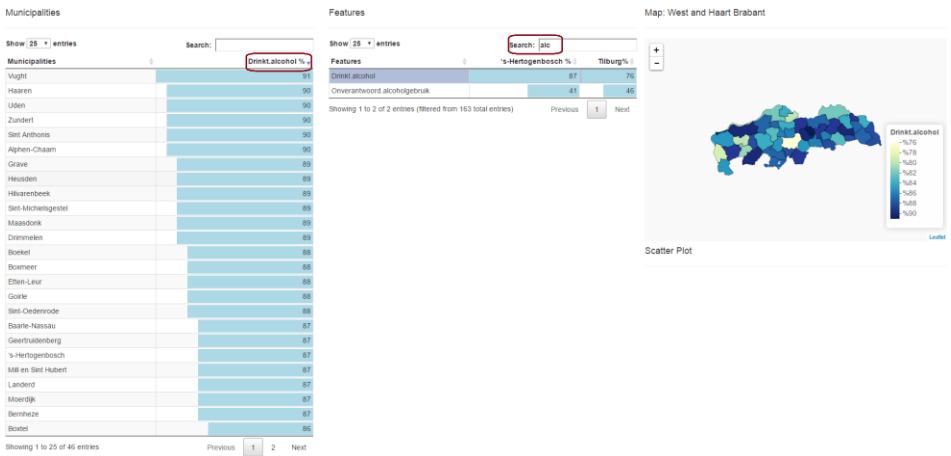

Figure 5.9: Comparing the alcohol consumption percentages in different municipalities.

Finally we want to see if there is any relation between the alcohol consumption and obesity rates. At the top right search bar we type '**obe**' and we select the term Obesitas. Keep in mind that right now two different features have been selected, and for that reason the scatter plot is created. The map is updated and shows the obesity rates (Figure 5.10). The scatter plot shows the obesity rates against the alcohol consumption.

Figure 5.10: Comparing the alcohol consumption and obesity rates in different municipalities.

It is not clear in the scatter plot if there is any relationship between the characteristics. We can only see a cloud of scattered points at the bottom right of the scatter plot. In order to zoom in, we just select the region of the scatter plot we would like to focus on. We then obtain the scatter plot that is shown on Figure 5.11. Notice that every point corresponds to a different municipality. There is a trend at this scatter plot: A higher alcohol consumption correlates with lower obesity rates.

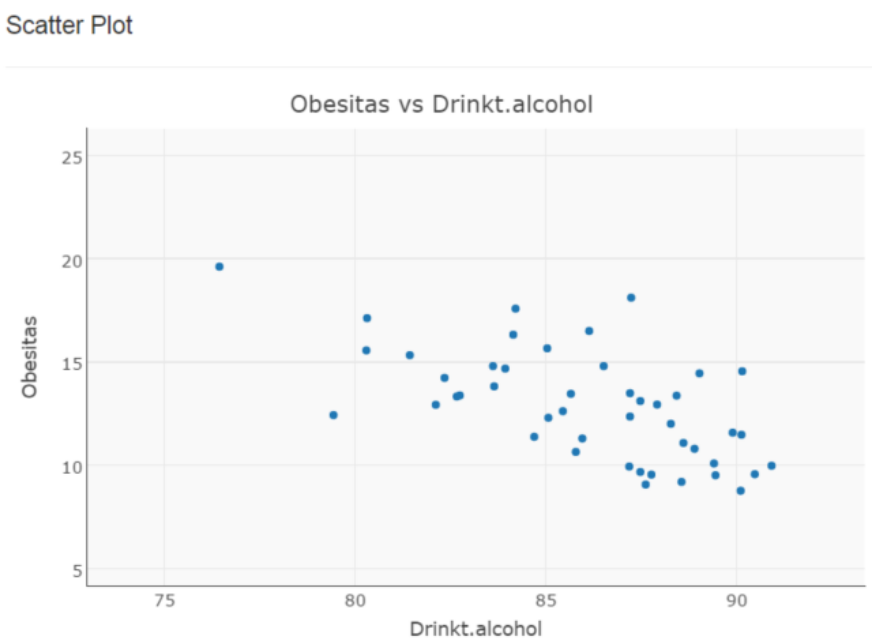

Figure 11: Obesity rates against alcohol consumption.

6. Conclusions

I^f-F^I

We presented a new method for the exploration of high-dimensional multivariate data using two tables with simple, well-known interaction techniques. The method is generic and has successfully been tested with a mix of numerical and categorical multivariate data. The minimalist approach has its limitations, but we argue that it would be a useful addition to existing multivariate data visualizations. This visualization technique allows the user to explore the data in a user friendly way such that the goals as presented in Section… are tackled [2].

Scatter Plot

We extended the I^f-F^I tables' functionality with the scatter plot to facilitate the exploration of correlations among different features. The scatter plot presents a simple and powerful visualization tool that helps to study and observe possible patterns in the data set.

Map

We introduced a new functionality to the I^f-F^I table that facilitates the visualization of the features for every municipality. Making comparisons between different regions is usually not efficient just by observing percentages in a table. This map facilitates the exploration of geographic effects.

In general we conclude that:

- A visualization tool that combines I^f-F^I table with map and scatterplot functionalities facilitates the exploration of multivariate monitor data from different perspectives.
- The tool is interactive and easy to understand and compare municipalities for a given feature.

- The tool also provides the functionality of exploring the relation and pattern between two features.

7. Recommendations

We recommend that GGDs and its end-users use the visualization tool to explore the monitor data. The tool is easy to use and interactive to find insight in the multivariate monitor data.

Future Work

We integrated several features in our visualization application. Of course, there are some steps to be taken. The developed tool visualizes the dataset that was obtained in the year 2012. Since there is a new dataset generated in the year 2016, it will be interesting to explore this dataset in a similar way. In this project two different GGD communities are involved. Our visualization tool includes data for regions that belong to both GGD communities and for that reason the visualization should be updated by showing these regions separately. Finally, the visualization is made for studying features on municipality level. In many cases, the end users are interested to study in detail what happens on district level, and this should also be taken into account for future work.

References

[1] “Exploring Items and Features with I^f,F^l-Tables”, Corput, Paul van der; Wijk, Jarke J. van, Computer Graphics Forum, 2016, volume: 35, issue: 3, startpage: 31

[2] “A nested Model for Visualization Design and Validation”, Tamara Munzner, IEEE Transactions on Visualization and Computer Graphics, Volume: 15, Issue: 6, November 2009, Pages: 921 - 928

Appendix

Used Software

For this software we used the R programming language and RStudio. We also used the following libraries:
Ggplot2, maptools , rgeos, Cairo, ggmap, scales, RColorBrewer, dplyr, devtools, rgdal, leaflet, Shiny, htmltools, htmlwidgets, plotly, DT.