Master Applied Data Science

Area comparisons on municipality, neighbourhood, and borough level: Shiny App in R for open data about amenities and health

Author:
Imke Dekkers (9690182)

Supervisors:
Laura Boeschoten
Erik-Jan van Kesteren

July 1, 2022

Abstract

Open government data can contribute to more transparency and a participatory governance. If the data is neighbourhood level data, it can support community-led actions and lead to changes in communities. However, open data can sometimes be difficult to find and interpret. This paper presents a Shiny application for the visualization of open data about amenities and health so that insights can be offered on both the state of the area and the comparison with other areas such that the average citizen is able to make sense of it independently. This is done by using the **shinydashboard** package to make a visually attractive app which includes two different tabs: for the data about amenities and about health. The app appears to be comprehensible for the average citizen, but for people without data literacy skills, data intermediaries may be necessary.

Keywords: open government; open data; amenities data; health data; Shiny app;

1 Introduction

In recent years, an increasing amount of open government data movements have emerged, an example is the establishment of The Open Government Partnership (OGP) in 2011. This partnership united government leaders and civil society advocates to foster transparent, participatory, inclusive, and accountable governance. 77 countries and 106 local governments that represent more than two billion people are involved in this OGP (OGP, 2022), illustrating the importance of the open government data movements.

An open government allows citizens to monitor and influence government processes. Governments can be more open by improving their citizens' access to information and access to decision-making arenas. In order to have that access to information, open government data is necessary, which is data that is made available by the government without restrictions and is easily found and accessible (Veljković et al., 2014). Publishing them has various benefits, including more transparency, releasing social and commercial value, and a participatory governance (Attard et al., 2015). By having open data, the information asymmetry between government and citizens decreases, which leads to a reduction in corruption (Meijer et al., 2012). Moreover, citizens are able to observe government initiatives and their decision-making and can confirm whether or not conclusions drawn from the statistics are accurate and justified. This in turn leads to the possibility of active participation of citizens in government decisions and policy making which allows the government to become more citizen-centered.

These are mainly general benefits of open data. However, many people are primarily interested in how they could benefit from open data on a more individual level. Access to neighbourhood level data and decision-making arenas can support community-led actions and lead to changes in local communities (Yoon & Copeland, 2019), since community members often have more knowledge and a better understanding of their community than policy makers. Therefore, policy makers do not always make the best choices for the community. If data about local areas is being shared, this leads to a shift from conversation-based to evidence-based conversations in communities. Policy makers and community members now also have the same information available to them and can work together to identify or solve local problems.

While there are many potential benefits to open data, there are also a number of challenges that need to be overcome before open data can become beneficial. Janssen et al. (2012) provide a list of barriers, which are divided into the following categories: institutional, task complexity, use and participation, legislation, information quality and technical. These are either barriers for data providers or data users, but the focus in this paper is on barriers for data users, which can mainly be divided into task complexity, and use and participation. A big problem when data is being published, is that is too difficult for the average user. It can often be a challenge to find the appropriate dataset, to understand its meaning, what the use of it could be, and they are often too complex such that it is difficult to search and browse to find the right information in the data. Moreover, the average user may not have enough (statistical) knowledge to make sense of the data. Many initiatives exist that try to visualize data in a comprehensible manner, but the creation of too many initiatives also leads to frustration of the user (Janssen et al., 2012).

Before trying to visualize the open data, the needs of the citizens should be carefully considered by involving them in the conversation. For example in Overvecht (municipality of Utrecht), the municipality started the project "Samen voor Overvecht" together with residents and professionals with the improvement of the neighbourhood as goal (Samen voor Overvecht, 2022). To do this, inspiration sessions have been organized where residents of Overvecht have formulated multiple information requests. A recurring theme in these questions is about the amount of and distance to various amenities. For example, one question they had was how many restaurants there are in Overvecht (Wegdam, 2022b). Publishing data about amenities on a local level could have multiple benefits for citizens and government agencies. Municipalities often try to involve citizens in their decision making, but citizens sometimes still feel that their complaints are being disregarded without any offered explanation (Helden et al., 2009), an example is when residents disagree with the build of new facilities. Open

data about the proximity to amenities per neighbourhood could facilitate transparent decision-making such that citizens can understand and accept the decision better (Ruijer et al., 2017). But this type of open data could also be beneficial the other way around. When citizens think that the distance to a certain type of amenity is too long, this data can help to either confirm or deny these claims by comparing data of a local area with other areas all over the Netherlands. If their suspicions are then proven to be correct, this data can help them make a strong case for the municipality to change this.

Another theme that was raised by the residents of Overvecht is health. In one inspiration session they mention that biking and walking should be given more attention (Wegdam, 2022a). Moreover, they note that their health is low compared to other neighbourhoods in Utrecht (Ruijer & Dymanus, 2022). In the "Preventieakkoord", multiple goals are stated for which the Dutch government is working together with social organisations since 2018 to make people healthier (Rijksoverheid, 2018). To reach these goals, actions are being taken on a national level, but also on a local level since research has shown that a local approach would be beneficial (Wesselink, 2020). Local health data can play an important role in this, since it gives the opportunity to identify where the health related problems are the biggest (Overvecht for example) and which areas should thus be given the most attention. It would also be beneficial if the local areas can be compared with areas all over the Netherlands, to identify areas with similar problems, who could perhaps learn from each other.

Local data about amenities and health in the Netherlands are already being published. CBS (Centraal Bureau voor Statistiek) (CBS, 2022) has released data about amenities and RIVM (Rijksinstituut voor Volksgezondheid en Milieu) (RIVM, 2021) about health. However, these datasets are not easy to use for most people because they are too complex to handle and too difficult to make sense of. To be able to start a continuous conversation between a government and their citizens, it is necessary to go beyond the provision of access to data. Tools or instruments should be implemented that citizens can use to understand and interpret the data such that they can derive valuable insights (Janssen et al., 2012).

Various tools for the visualization of open data about amenities and health are already in existence. Many municipalities have created dashboards that include various topics, an example is "Utrecht in Cijfers" (Gemeente Utrecht, 2022), where you can compare neighbourhoods within the municipality with each other. With "Waar Staat Je Gemeente" (VNG, 2022), it is possible to compare your municipality with another one or with all municipalities which have the same level of urbanity. For health related data, RIVM has published a tool (RIVM, n.d.) where a map is shown per variable with the values for all municipalities in the Netherlands. When the user then clicks on a municipality, all neighbourhoods within this municipality are shown. One thing that is missing in all of these tools, is the possibility to compare neighbourhoods all over the Netherlands with each other.

A tool is desired where you can switch between different levels of areas and compare them with areas across the Netherlands. Therefore, the following research question has been formulated: Can open data about amenities and health be visualized so that insights can be offered on both the state of the area and the comparison with other areas such that the average citizen is able to make sense of it independently? To answer this research question, a Shiny app has been developed in R.

The remainder of this paper is structured as follows: first, the data that has been used in this research is described. Then, a detailed description of the visualisation tool follows where it is explained what elements the app consists of and why. Subsequently, two application examples are presented where examples of how the app can be used are shown. Finally, a discussion and conclusion are given, including the limitations of this research and opportunities for future research.

2 Data

The data that has been used in this research is all openly available data from either the Dutch "Centraal Bureau voor Statistiek" (CBS) or from "Rijksinstituut voor Volksgezondheid en Milieu" (RIVM). From CBS a dataset about key figures and amenities (CBS, 2020b) (called amenities from now on) and a dataset containing zip code information (CBS, 2020a) (called zip codes) were used. From RIVM, a dataset about the health per area (RIVM, 2022) (called health from now on) was used. All datasets consist of data about the year 2020 (except for the health data which also contains data about 2012 and 2016) and use the municipal division of 2020. They all contain information on three different levels: the first level is the municipality ("gemeente"), this is divided into neighbourhoods ("wijken") which is the second level and these are divided even further into boroughs ("buurten"), the third level.

In the remainder of this section, the three used datasets are explained in more detail, and the ethical and legal implications of the data are discussed.

2.1 Amenities data CBS

The amenities dataset contains a wide range of information, for example about the population, age distribution, migration backgrounds, houses, and geometries. Additionally, data about the proximity to different types of amenities is included. This is the data that has primarily been used from this dataset (see Table 2 in the appendix for all used variables). The amenities have been divided into eight categories:

- Health and well-being
- Retail
- Hospitality industry
- Childcare
- Education
- Green areas
- Traffic and transport
- Leisure and culture

Unfortunately, all data about the green areas amenities were missing, so only seven of the categories remained. The amenities data consists of two types of variables. The first type is the average distance people within a certain area have to travel to for example a hospital. For the calculation of the distance, only paved roads that are accessible for cars have been used and the average distance has only been recorded for areas with at least 10 inhabitants. Moreover, it is only recorded if for 90% or more of the inhabitants of the area the exact location of their address could be determined. The second type of variable for the amenities data is only available for some of the amenities and consists of the amount of amenities within three different radii. For the calculation of the average amount of amenities within a radius, again only paved roads are taken into account (CBS, 2021).

Before this dataset could be used for the Shiny app, some pre-processing steps had to be done. All pre-processing steps have been applied to the three-level tables. Aside from data about the municipalities, neighbourhoods and boroughs, the data contained information about areas with only water. These areas could be identified with the variable called *H2O*. When an area only consists of water, this variable has as value "Ja" and otherwise the value is "Nee". The data about these areas were almost all missing since there are no people living there. As this data was considered to not be of use in this research, the rows with "Ja" as value for the variable *H2O* were discarded. The remaining data consisted of 355 municipalities, 3177 neighbourhoods and 1380 boroughs.

The missing data in the dataset were made clear by setting those values to -99999999. This means that these values are either unknown, insufficiently reliable or secret. However, R does not recognize this as missing data and treats these as actual values. Therefore, all values of -99999999 were changed to NA. A detailed table of the amount of missing data can be found in the appendix in Table 2.

Furthermore, a lot of variables are not used for this research, so these variables were discarded. Table 2 in the appendix also shows what variables have been used. The dataset contained multiple variables about income such as the amount of people with an income and the mean income per resident. However, in the downloaded dataset these variables only had zero values. The table from database StatLine of CBS "Kerncijfers wijken en buurten 2020" showed that the income variables were not all zero at all (CBS, 2022). Therefore, this data was downloaded separately as this was thought to be necessary data for this study. These variables were joined with the amenities dataset by the municipality, neighbourhood and borough codes. Additionally, to be able to compare areas based on a similar income, the areas were divided into four groups. To do this, the variable about the number of households below or around the social minimum was used and the areas were divided into four equally sized groups based on this variable.

The dataset is an ESRI Shapefile and contains geometries with information about where exactly the areas are located, which is important for visualizing the areas. The creation of the maps has been done with the **leaflet** package, which expects the spatial data to be specified in latitude and longitude using the coordinate reference system (CRS) WGS 84 (Cheng et al., 2022). Therefore, the data has been transformed to this CRS by using `st_transform()`. However, loading the geometries takes a long time, which is why the geometries were simplified with the `ms_simplify()` function from the R package **rmapshaper** (Teucher & Russell, 2021). As a result, loading the maps goes faster. The centroids of the different areas have been calculated in the pre-processing steps and have been added to the dataset. For the calculation of the centroids, the `st_centroid()` function from the **sf** package (Pebesma, 2018) has been used. This step does now not have to be taken in the app itself, which means that this speeds up the app.

2.2 Zip codes data CBS

The zip codes dataset consists of multiple tables. One table contained all PC6 codes together with their municipality, neighbourhood and borough code and the information about their names could be found in three separate tables. To have all information in one dataset, the table with the PC6 codes was joined with the three other tables. Now, there is one dataset containing all unique PC6 codes in the Netherlands, and the codes and names of what municipality, neighbourhood and borough they are located in. It is possible for a PC6 code to fall in two boroughs. In that case, there are two rows for this PC6 code, both with a different borough name.

2.3 Health data RIVM

The health related dataset used in this research has been obtained from the RIVM. This data is derived from questionnaires that have been filled out by over 540.000 people of 18 years and older for the year 2020. Even though a lot of data has been accumulated this way, there are not enough respondents to use weighing methods to calculate the numbers for all neighbourhoods and boroughs in the Netherlands. Therefore, RIVM developed a model to calculate these numbers based on the surveys. This is called the SWAP-model (SMall Area estimates for Policymakers).

The questionnaire from the RIVM contained questions about subjects that can be divided into three categories:

- Health and disabilities
- Lifestyle
- Participation and environment

The respondents were anonymously linked in a secure environment to CBS registration files which include information about background characteristics such as age, income and education. The RIVM used a XGBoost model to relate the health and lifestyle data to these background characteristics. XGBoost models are broadly used in machine learning problems and is a scalable end-to-end tree boosting system (Chen & Guestrin, 2016). The spatial location has also been taken into account in this model. The expected health and lifestyle of all adults have been calculated this way and the results have been averaged over the regarding neighbourhood or borough. For areas with less than 10 residents, the data is not shown (RIVM, n.d.).

The dataset not only contains data about the year 2020, but also from the years 2012 and 2016. As there have been some municipal reorganizations since then, these numbers have been recalculated for the municipal division of 2020 by the RIVM. Therefore, the data is comparable over the years. The questionnaires from the different years were not completely the same however, so some variables are not available for the years 2012 and 2016. There are also different age classes available, all variables have three different values for the varying age classes: 18-65, 65 and older, and 18 and older.

This dataset has been joined with the amenities dataset from CBS to get the geometries and other information such as age distribution and income. The join has been done by the codes of the municipalities, neighbourhoods and boroughs. In the appendix in Table 3 to 5, the used variables and the amount of missing data can be seen, and Table 6 shows a description of the used health variables.

2.4 Ethical and legal considerations

All of these datasets do not contain any personal information. The data cannot be traced back to individuals.

A condition for the trust of citizens and obtaining better effects from open data initiatives is good quality of the released data. Key components of data quality are timeliness, availability of metadata, accuracy, and usefulness. (Safarov et al., 2017). All of these components are considered to be of good quality. Only for the health data on neighbourhood and borough level, the data accuracy for neighbourhoods and boroughs can be questioned as these are estimates. With the XGBoost model from the RIVM, reality is approached as closely as possible, but the data remain estimates of reality. This data should thus be treated with caution. Overall though, the data quality is very high for this research. Therefore, it can hopefully contribute to the trust of citizens in the data.

3 Visualization tool

To answer the research question, a visualization tool has been developed. A dashboard has been created in R with the **Shiny** package (Chang et al., 2021) which makes it easier to build interactive web apps. A Shiny application consists of two important components: the server and the user interface (UI). The sever defines how

the app works, while the UI mostly defines how the app looks and specifies where the input fields and outputs such as visualizations are placed. The input is sent from the UI side to the server side, which then does its calculations and makes visualizations. These are in turn sent back to the UI such that it can display them (Wickham, 2021).

For this research, two categories of open data have been visualized: data about amenities, and data about health. These two categories both have a separate tab in the app. To get interim feedback on the visualization tool, an earlier version of the Shiny app has been shown to a variety of people, including visitors of the "Dag van de Buurt Overvecht". This was a day in the neighbourhood Overvecht (municipality of Utrecht) where multiple activities were organized, including an information market. Here the tab for amenities was demonstrated for people who were interested, to collect feedback on how comprehensible the app was for them.

The remainder of this section describes the app, including how the app looks like and what features it consists of (taking the on the "Dag van de Buurt Overvecht" collected feedback into account). First, the tab for the amenities data is shown and explained, followed by the tab for the health data. Finally, some general adjustments for an increasing user-friendliness that apply to both the amenities and health tab are given.

3.1 Amenities

For the amenities tab, first the graphical user interface is explained. Then some elements of the app and how they are generated are described in more detail. The goal of this tab is to visualize the data about amenities in an insightful an comprehensible manner.

3.1.1 Graphical User Interface

To build the graphical user interface (GUI), The **shinydashboard** package (Chang & Borges Ribeiro, 2021) has been used within the UI part of Shiny to create an attractive dashboard. In Figure 1 an overview of the GUI can be found. On the left, it can be seen that the first tab has been selected which is for the amenities. For the second tab (health), see section 3.2. For the tabs about traffic accidents and crime, see the papers of Wooning (2022) and Kellij (2022) respectively.

Once the app is opened, the user only sees the parts where something needs to be filled in, but no actual plots (so only the explanation box (1), the zip code searcher (2), the area selection box (3), and the theme selection box (7) from Figure 1). The reason being that it can be too overwhelming if everything is loaded at once and the user might not know where to look. Now, the user can first focus on filling in the right information to get the desired visualizations. One of the collected points of feedback on the "Dag van de Buurt Overvecht" was that it is not immediately clear what they are looking at. By loading the different plots sequentially, the hope is that this problem is partly solved.

Another related point of feedback was that it was a bit difficult to understand what users needed to do. Therefore, an explanation box (1) was added to the top left corner of the tab. By clicking on the '+' sign, a detailed explanation of the app and how it works appears (see Figure 2). The user can read this first so they know how the app works. In this description, it is also explained that if they do not know the name of their neighbourhood or borough, they can use the box zip code searcher box (2 in Figure 1) (see section 3.1.2 for more information).

After that, the box to select the area (3) can be filled in. First, the area level needs to be selected. Then, the desired area should be chosen (see section 3.1.3). And the last thing that needs to be filled in within this box is the "Vergelijken met" input field. Here, the user can choose the areas that they want to compare their selected area with (see section 3.1.4). Finally, the "Zoeken" button needs to be pressed such that more boxes with information/maps are shown (information box (4), box with map (5), and 5 similar areas (6)).

In the information box (4), the urbanity and income level are given together with some explanation of what these levels mean. The user now knows some information about their selected area. At the "Dag van de Buurt Overvecht", one person mentioned that it was not immediately clear what the information in this box was about. Therefore, the title of the box now includes the selected area name such that it is clear that this information is about the selected area.

Next to it, there is a box with a map (5). Here the user can see what the location of their selected area is (blue pointer in the map) and what the comparable areas are (blue areas). The map also includes five red pointers to other areas than the selected area. These are the five areas that are mentioned in the adjacent 5 most similar areas box (6). Here, the five most similar areas as the selected area based on the amenities data are given (in section 3.1.5 a more detailed explanation is given).

Once the user has absorbed all of this information, they can scroll down to the themes selection box (7) where a user needs to select what kind of variable they are interested in. First, the theme can be chosen. There are seven possible themes: health and well-being, retail, hospitality industry, childcare, education, traffic and

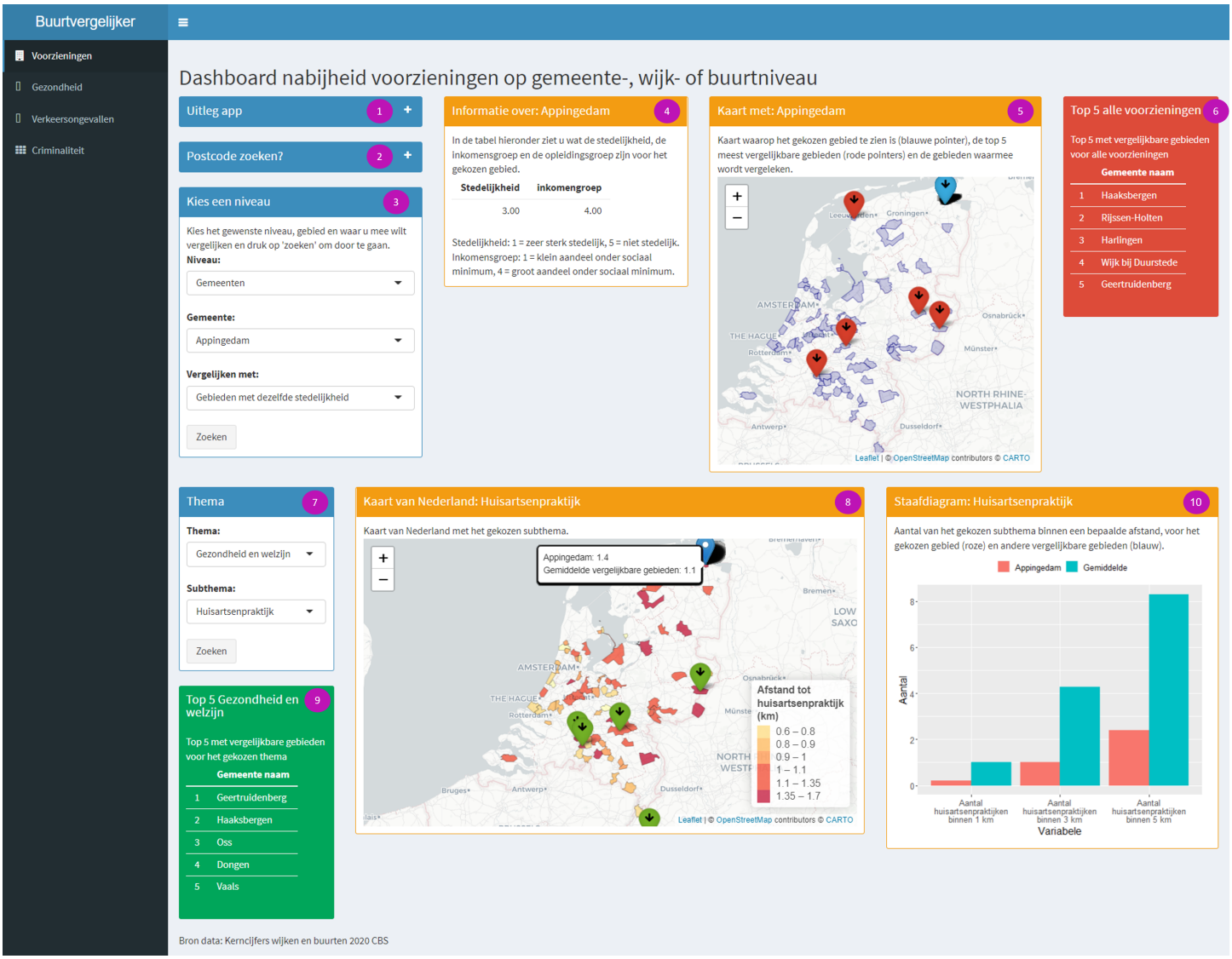


Figure 1: Graphical user interface for amenities tab: (1) Box with explanation of the app, (2) Box to find area name based on zip code, (3) Box to select level and area, (4) Box with information about selected area, (5) Box with map showing selected area, comparable areas, and the 5 most similar areas, (6) Box with 5 most similar areas, (7) Box to select theme and subtheme, (8) Box with colored map based on subtheme, (9) Box with 5 most similar areas based on selected theme, (10) Box with barplot of selected subtheme.

transport, and leisure and culture. These are further divided into subthemes where the user can choose from. For example for the health and well-being theme, the subthemes are: general practice, hospital, and pharmacy. When the theme and subtheme have been selected, the "Zoeken" button can be pressed, after which additional visualisations appear (box with map of subtheme (8), 5 most similar areas box (9), and barplot box (10)).

In the map of subtheme box (box 8), the map shows the distance to the amenity that has been selected as subtheme. Again, the blue pointer shows the selected area. The green pointers are the most similar areas based on the amenities data of the selected theme that are mentioned in the green box on the left (9 in Figure 1). A more detailed explanation on the 5 most similar areas and the map can be found in section 3.1.5 and 3.1.7 respectively.

On the right, the barplot box (10) can be found. This shows a barplot with the amount of amenities (of the selected subtheme) within various radii. Not all subthemes have data available on the amount of amenities within a certain radius. If such a subtheme is selected, this box will simply not be displayed. More information of this barplot can be found in section 3.1.8.

Finally, in the left corner at the end, the source of the data is mentioned to be transparent about where the data comes from (in this case from CBS).

Figure 2: Explanation box for the amenities tab.

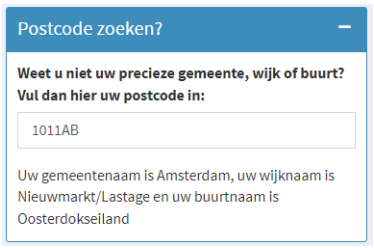


Figure 3: Box for finding area names based on zip code.

3.1.2 Zip code searcher

Most people know what municipality they live in. However, a lot of people do not know the name of their neighbourhood or borough as was noticed on the “Dag van de Buurt Overvecht”. This is a problem, since the user then does not know what to fill in as neighbourhood/borough name. Therefore, the zip code searcher has been added (see Figure 3). It is assumed that most people do know what their zip code is. The user can simply give their zip code and the app will give the municipality, neighbourhood, and borough name. This is done in the server side of the app by first removing any spaces in the given zip code as the used dataset does not contain any. The by the user given zip code is the input for the function `str_replace_all()`, where any possible spaces in the given zip code are replaced by an empty string. Lower case letters are changed into upper case letters since that is the way the zip codes are recorded in the dataset. This is done by the function `toupper()`. Then, the app searches for a match between the given zip code and the ones in the zip code dataset. The output will be the municipality, neighbourhood, and borough name of all matches since it is possible that a zip code occurs in two boroughs. With this feature, the user is able to find out their area names.

3.1.3 Selection area level and area

Users of the app might be interested in various area levels. One user might want to know more about their municipality while another wants to look at their neighbourhood. Therefore, it is possible to select the desired area level. There are three possible choices: municipality, neighbourhood, and borough. Depending on the user input for the area level, either one, two or three input boxes are shown to select the desired area in the area selection box. To facilitate this, the `conditionalPanel()` function has been used three times in the UI side of the app. The first conditional panel is only shown when the selected area level is municipality. In this case, one input box is shown where the user can select their desired municipality. The choices consist of all unique municipality names, which are determined with the `unique()` function. The second conditional panel is only shown when the selected area level is neighbourhood. Two input boxes appear in this case: one for the municipality and one for the neighbourhood. The choices for the municipality are again all unique municipality names. The possible choices for the neighbourhood depend on the selected municipality. Only the neighbourhoods in the selected municipality can be chosen. To determine what these neighbourhoods are, the `observeEvent()` function has been used in the server side of the app. This function observes the municipality input and uses this for the `updateSelectInput()` function to select only the neighbourhoods that have the same municipality name as the selected municipality. The third conditional panel is shown when the selected

area level is borough. This panel has three input boxes: for the municipality, neighbourhood, and borough. The choices for the municipality and neighbourhood are determined in the exact same way as before. The choices for the borough consists of all boroughs in the selected neighbourhood. What these boroughs are is determined in the same way as for the neighbourhoods.

3.1.4 Finding the comparable areas

In the app, values of the selected area are being compared with the average multiple times. But what areas should be taken into account when calculating the average value? This depends on the wishes of the user. Someone might want to compare with all other areas in the Netherlands when they are just interested in their position relative to the Netherlands. For another user who wants to know whether it is normal that he has to travel very far to for example the hospital, it might make more sense to compare with areas that have the same urbanity level, since it is logical that you need to travel longer to amenities when you live in a rural area than when you would live in the city. Therefore, it has been made possible to choose with what kind of areas you want to compare. There are three possible choices: all other same level areas in the Netherlands and areas with the same urbanity level as mentioned earlier. Additionally, the user can also choose to compare with areas based on the same income level, as it could be the case that areas with a higher income level have more amenities than areas with a lower income level. The urbanity level is one of the variables in the amenities dataset from CBS, where level 1 means that it is very urban and level 5 is not urban at all. The income level has been calculated in this research. Multiple variables about income are present in the amenities dataset. However, for neighbourhoods and boroughs these variables are missing to varying degrees. The choice has been made to use the variable with the least amount of missing data, which is the number of households beneath or under the social minimum. This variable is missing in 0% of the municipalities, 8.59% of the neighbourhoods and 31.21% of the boroughs. Since this is a very high percentage for the boroughs, the choice was made to disable the income option for boroughs. So on this level it is only possible to compare with all areas in the Netherlands or with areas with the same urbanity level.

Every time the user selects another area level, area, or areas to compare with, the app updates the used dataset for all the visualizations. To do this, a function inside the server side of the app has been created to filter the dataset. Figure 4 shows what this function does.

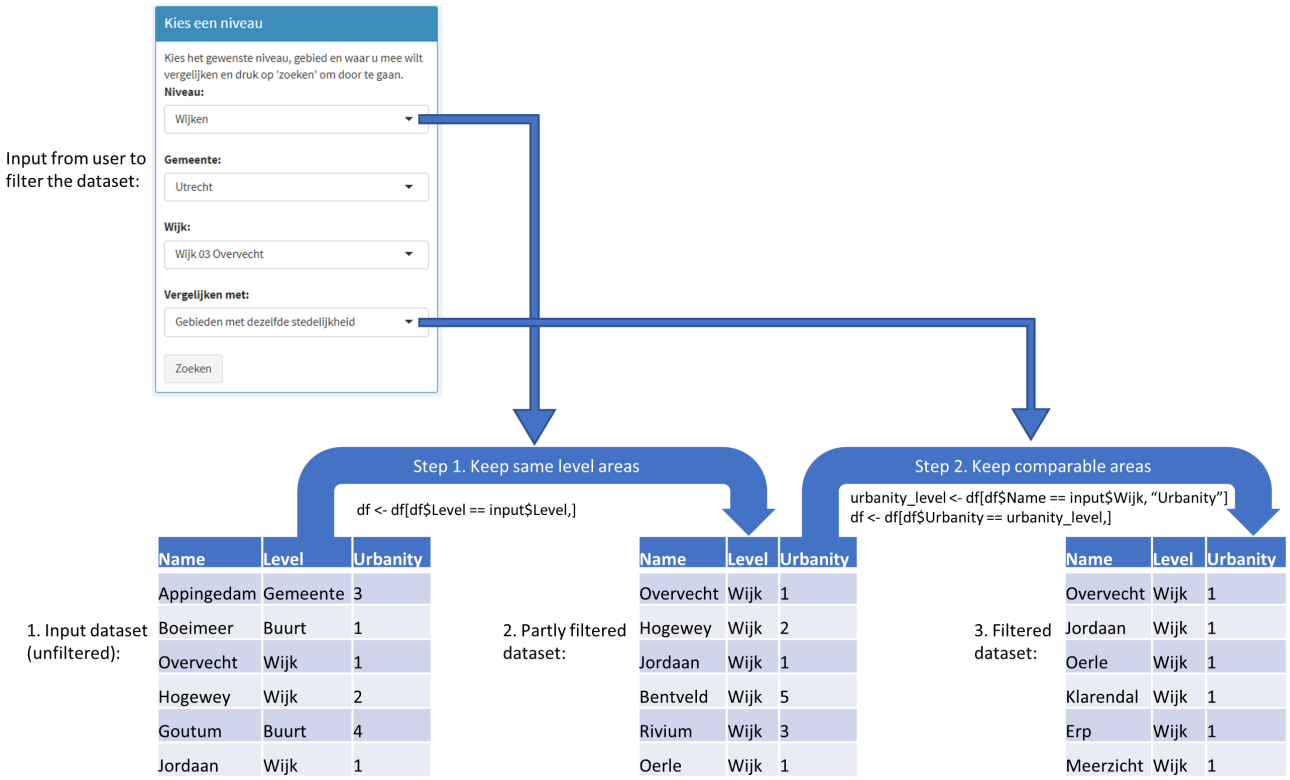


Figure 4: Steps in function that filters the dataset. Input for the function is the user input and the amenities dataset. The areas with the desired level and comparability are kept to obtain a filtered dataset.

The input for the function is the amenities dataset and the input from the user. Step 1 is to filter out the

areas that do not have the same level as the user has selected. Step 2 is to keep only the areas in the dataset that are going to be compared with the selected area. If the option to compare with all areas in the Netherlands is selected, no rows are discarded in this step. When the option of comparing with the areas with either the same urbanity level (like in Figure 4) or the same income level is selected, first this level of the selected area is determined. Then, all rows that do not have this same level, are discarded. If this level of the selected area is missing, no rows will be discarded either and the selected area will be compared with all same level areas in the Netherlands instead. To make this clear to the user, this will be displayed in a red warning message.

The function will now return a filtered dataset containing the selected area and all other areas that the selected area is going to be compared with. This dataset can be used for the other visualizations.

3.1.5 Finding the top 5 most similar areas

Another feature of this app is the ability to find areas that are similar to the selected area based on the data about amenities. This is relevant for users who have detected a certain problem in their area and want to know how other areas deal with similar problems. This way, they can learn from each other and try to solve their issues. Therefore, the app gives the five most similar areas based on all amenities variables (box 6 in Figure 1) and the five most similar areas based on all amenities variables in the chosen subtheme (box 9 in Figure 1). To find these areas, multiple steps are taken. First, the filtered data has to be obtained (explained in section 3.1.4). Then, the data about the amenities are scaled by calculating the Z-score. This is necessary because the amenities variables have different scales. The formula of calculating the Z-score is:

$$Z = \frac{(\chi - \mu)}{\sigma}$$

where χ = the observed measurement, μ = the population mean, and σ = the population standard deviation. The function `scale()` is applied to the data, which uses this same formula. The result is rescaled data with mean 0 and a standard deviation 1. An individual rescaled value describes how many standard deviations the observed measurement is away from the mean. With a positive score indicating it is above the mean and a negative score indicating it is below the mean (Curtis et al., 2016). Then, the distance is calculated from the selected area to all other comparable areas with the `dist()` function. Multiple distance metrics are possible, but the choice has been made to use the Manhattan distance metric, since this metric is preferable when the data has a high dimensionality (Aggarwal et al., 2001). As there are more than 80 variables on the amenities, this is definitely the case. The Manhattan distance d between point x and y can be calculated as follows:

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

After these distances are calculated, the dataframe is ordered by this distance with the function `order()`. Then, the top five areas are returned to get the areas with the smallest distance to the selected area. The output is the area names of these five areas. For some areas, no data on the amenities is available. In these cases, a message will be returned saying that there is insufficient data for the selected area.

3.1.6 Map with selected area

In the app, a comparison is made with the selected area and comparable areas multiple times. Therefore, it would be desirable if the user can also discover what these comparable areas are and what their location is. Moreover, it would be useful if the user can see where the top 5 most similar areas are located. Thus, a map has been created with the location of the selected area (blue pointer), the location of the 5 most similar areas (red pointers) and the location of the comparable areas (blue areas) (box 5 in Figure 1). For this map, the **leaflet** package (Cheng et al., 2022) has been used. The input data for this map is the filtered dataset that has been explained in section 3.1.4. As sometimes not all areas in the Netherlands are shown on the map when the user wants to compare with similar income or urbanity areas, the map may contain a lot of gaps. It can in this case be hard to determine where exactly in the Netherlands the areas are located. Therefore, a background map is used as a reference by using the `addProviderTiles()` function from **leaflet**. The used provider is *CartoDB.Positron*. With the `addPolygons()` function, the areas in the filtered dataset are displayed on the map. With the help of the `label` option within this function, the names of the areas appear when the user hovers over them. For the pointers, the earlier calculated centroid of the polygon has been used as input for the `addAwesomeMarkers()` function. A distinction between the selected area and the most similar areas has been made by giving the pointers a different color. The selected area has been given the color blue since this has been picked in a box which is also blue. Box 6 with the most similar areas is red, and therefore these areas are shown with red pointers. Leaflet also makes it possible by default to zoom in and out on specific areas.

3.1.7 Map with selected subtheme

It would be interesting for the user to see what the distance to the selected amenity is for both their selected area and the other comparable areas. Therefore, a map is shown were the areas are colored by the distance to the amenity (box 8 in Figure 1). Additionally, the green pointers show the 5 most similar areas based on the selected theme. Here the color green was chosen because some areas are displayed in red on the map and if the pointers would be red as well, they would not be as easily distinguishable as they are now.

The map has been created in a similar way as the map in the previous section. The main difference is that the polygons are not displayed in blue anymore, but are colored by the distance to the amenity that has been chosen as subtheme. This has been done by using the `fillColor` option within the `addPolygons()` function. There are two used methods for determining the colors of the areas. The default method is to use the `colorQuantile()` function, which slices the values of the selected subtheme into subsets with an equal number of observations. As palette, "YlOrRd" has been used which gives higher values a red color and lower values a yellow color. This method makes sure that even if there is one very high outlier, the rest of the areas are not only displayed in yellow as such a map would not be very informative. However, this method fails if the 'breaks' are not unique, which happens when the quantile values are the same for some levels. Therefore, if this method gives an error, the `colorBin()` function is used instead. This function performs binning based on the values of the selected subtheme. Again, the "YlOrRd" palette has been used.

To make the value of the selected area immediately visible, a text box appears near this area with the value of the selected area and the average value. For this, the `label` option in the `addMarkers()` function has been used. The average is calculated by only taking the areas in the filtered dataset into account. When the user hovers over the areas, these values also pop up, which is done by using the `label` option in the `addPolygons()` function.

3.1.8 Barplot of selected subtheme

The last element in this tab for the amenities is the barplot for the selected subtheme (box 10 in Figure 1). Aside from the distance to the amenities, there is data on the amount of amenities within three different radii. These have been visualized in this barplot. For all three radii, the value for the selected value and the average is shown. To do this, first the average value of the selected subtheme is calculated for all three radii. Only the areas in the filtered dataset have been taken into account when calculating the average. Then, the values for the selected area are extracted from the dataset. These values for both the average and the selected area are then put together in one dataframe with an additional column indicating if the value is for the average or for the selected area. This dataframe is then given as the input for the `ggplot()` function of the **ggplot2** package (Wickham, 2016). With the `fill` option, it is specified that the bars for the average and selected area should have a different color. Finally, the `geom_col()` function has been added to `ggplot()` to get the bars.

3.2 Health

The second tab in this app is for the health dataset from the RIVM for which the aim is to visualize this data such that it is insightful and understandable. Again, first the graphical user interface is explained. Then, some elements are discussed in more detail.

3.2.1 Graphical User Interface

In Figure 5 an overview of the health tab can be found. When this tab is first opened, it looks very similar to the amenities tab since not all elements are loaded immediately. Only the explanation box (1), the area selection box (2), and theme selection box (6) are visible right away. The first box is the explanation box (1) with again a description of what to do and what can be seen (see Figure 6). Next, the area selection box (2) allows the user to select the level and area. This works exactly the same as with the amenities. The only difference being that here there is an extra choice in the "Vergelijken met" input box, namely the choice to compare with areas with a similar age distribution (also see section 3.2.2). Once the user is satisfied with their choices and has pressed the "Zoeken" button, three more boxes show up (information box (3), age distribution box (4), and map box (5)). The information box (3) is similar as in the amenities tab and contains information about the urbanity and income level. The age distribution box (4) shows the age distribution of the selected area. The box with a map (5) displays the selected area (blue pointer) and the comparable areas (blue areas). This map is generated similarly as the map in section 3.1.6.

Next, the user can again select a theme and subtheme in box 6. For the theme, there are three choices: health and disabilities, lifestyle, and participation and environment. These are again divided into subthemes, but some variables in this dataset are very closely related. For example the variables about underweight, normal

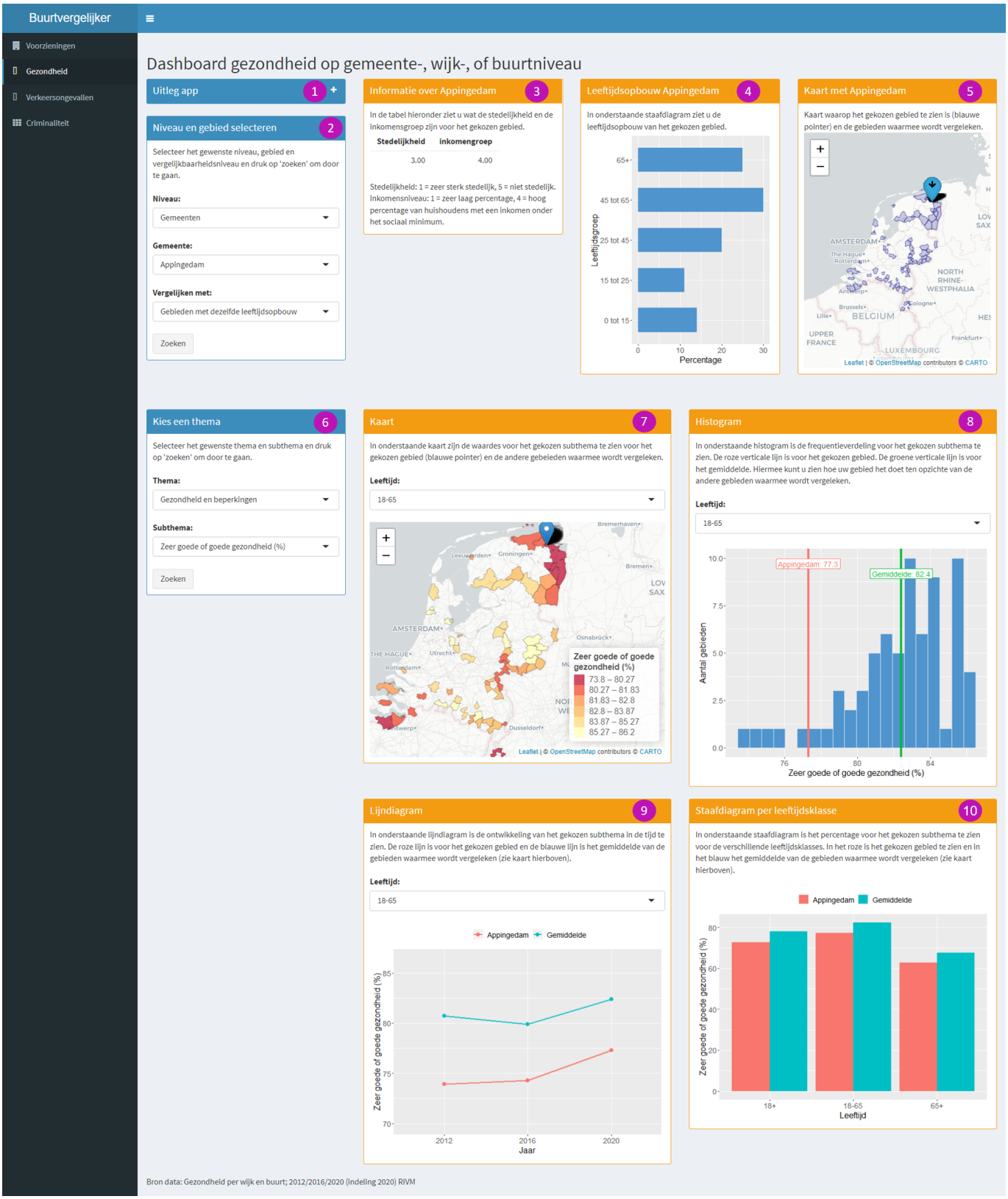


Figure 5: Graphical user interface for health tab: (1) Box with explanation of the app, (2) Box to select level and area, (3) Box with information about selected area, (4) Box with age distribution of selected area, (5) Box with map showing selected area and the comparable areas, (6) Box to select theme and subtheme, (7) Box with colored map based on subtheme, (8) Box with histogram based on selected subtheme, (9) Box with line chart of selected subtheme, (10) Box with barplot of selected subtheme for all age categories.

Figure 6: Explanation box for the health tab.

weight, obesity, and morbid obesity. These variables are therefore grouped together as one subtheme and are simply called "weight".

Depending on the selected subtheme, four or five additional plots show up. Normally, four different visualisations appear (a map, histogram, line chart, and barplot (box 7-10)). In the case that a subtheme has been selected that was grouped from multiple variables (such as the weight example), one extra visualization is shown first (see Figure 7). This is a barplot with the percentages of the different categories in the subtheme (see section 3.2.7). The other four visualizations are the same for all subthemes.

First, there is the subtheme map box (7) with a map where the areas are colored by the value of the selected subtheme (see section 3.2.4). Secondly, there is a box (8) with a histogram of the selected subtheme (see section 3.2.3). The third box is the line chart (9) (see section 3.2.5). For these three visualizations, there is an extra input box called "Leeftijd". With this input box, the user can choose from three age categories for which the plot is shown: 18-65, 65+, and 18+. Since the health of people can differ a lot for different ages, it makes more sense to look at the age categories 18-65 and 65+ separately. As the majority of the people are in age category 18-65, this has been made the default. If a user does not want to make a distinction between these age categories, it is still possible to see the data for 18+. However, this is the last option in the choice list.

The fourth visualization box shows the values of the selected theme for all three age classes (10) (see section 3.2.7).

If a subtheme such as weight has been selected (which contains multiple variables), an extra input box is given in these last four boxes called "Categorie". For example for the visualization with the map (see Figure 8). Here the user can choose for which variable the map should be shown.

Finally, in the left corner the data source is mentioned for transparency.

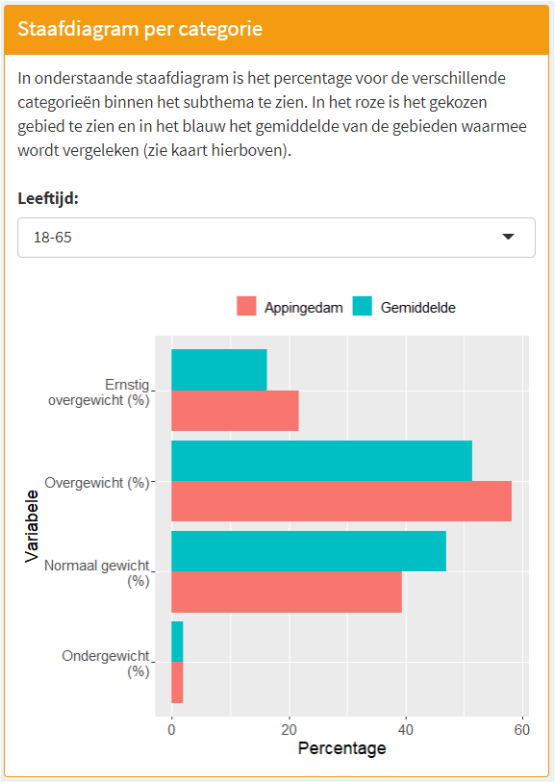


Figure 7: Barplot that shows the values of all categories in the selected subtheme (only shown when selected subtheme contains multiple variables).

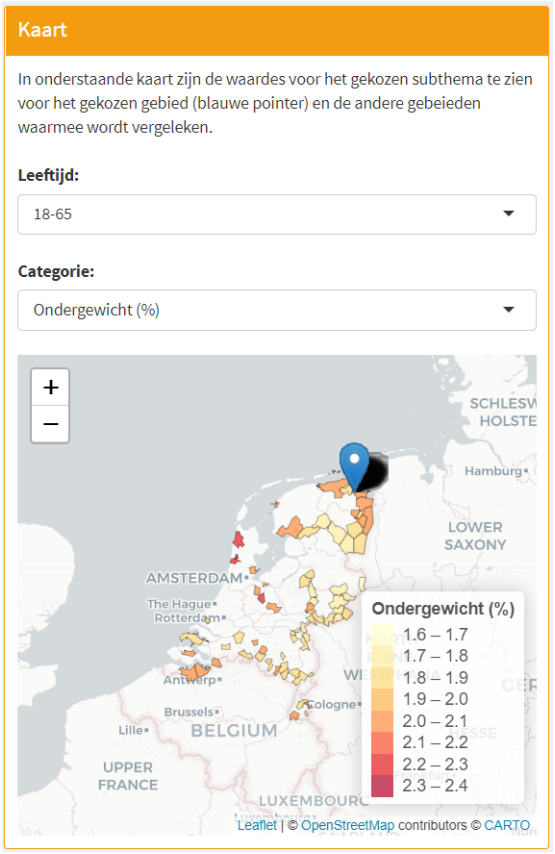


Figure 8: Box with extra input field "Categorie" in case the selected subtheme contains multiple variables.

3.2.2 Finding the comparable areas

The selection of comparable areas is done similarly as for amenities (see 3.1.4). However, the option to compare with areas with a similar age distribution is added, as the health of individuals is generally strongly related to their age. There are five variables on people's age in an area (see age distribution box (4) in Figure 5). Based on these five variables, the most similar areas are selected. This is done in a similar way as the top 5 most similar areas in the amenities tab (section 3.1.5). A dataframe with all same level areas is created (step 1 in Figure 4). This time, scaling of the features was not necessary as they are all already on the same scale. Instead, the Manhattan distance to all other areas is immediately calculated with the `dist()` function. When the distance is smaller or equal to ten, the area is considered to be comparable in age distribution. This number has been chosen because for municipalities 20% of all distances is 7 or lower and for neighbourhoods 20% of all distances is 12 or lower, 10 is thus chosen as middle ground. As there are a lot more neighbourhoods and boroughs than municipalities, enough comparable areas will remain. For areas with a very standard age distribution, more than 20% of the areas would be selected based on the cutoff distance of 10. Instead, it was chosen to only select the closest 20%. The reason for choosing this approach and not simply always selecting the closest 20%, is that there are some areas that are outliers based on their age distribution. For example, 96% of the residents of borough Hartelpark-West in Nissewaard are in the age category 15-25. Such areas will not have many similar areas based on the age distribution. Taking the closest 20% will then not make much sense as these will contain areas that are still not similar to the selected area at all. Thus, it was chosen to take a cutoff distance instead. Still, some areas are such outliers that there are less than five similar areas. Since it becomes very difficult to compare with such a low amount of areas, there is a minimum of five comparable areas. So, if there are less than five based on the cutoff distance, instead the five closest areas will be returned. If there is no data about the age distribution, all areas in the Netherlands are selected (see table 2 in the appendix for amount of missing data). This is again displayed in a red warning message.

3.2.3 Histogram of selected subtheme

The histogram (box 8 in Figure 5) shows the frequency distribution of the selected subtheme. In this histogram, only the areas that are in the filtered dataset have been taken into account (the areas that can be seen in the map box) and only the values of the chosen age category. The histogram has been generated with the `geom_histogram()` function with `bins=20`. A pink vertical line shows the value of the selected area. This way, it is immediately clear how the selected area performs relative to the comparable areas. The average value is also displayed in the histogram, with the green vertical line. These vertical lines are added to the histogram with the `geom_vline()` function, where the value of the selected area and the average is specified with the `xintercept` option. For clarity, the function `annotate()` has been added, to give the vertical lines a label that consist of either the selected area name or "Gemiddelde" (average) and their values.

3.2.4 Map with selected subtheme

The map (box 7 in Figure 5) also displays the values of the selected subtheme. This map has as additional benefit that you can also see what values belong to the comparable areas such that the spatial pattern becomes visible. The creation of the map is done similarly as the map in section 3.1.7. However, whether the color becomes more or less red the higher the percentage, depends on the type of variable. For example, having much stress is clearly a bad thing and in this case the higher the percentage, the more red the area is. But a high percentage of people who have a good health is for example a good thing. For such variables the lower the percentage, the more red the area. In these cases, the argument `reverse = TRUE` has been added to the `colorQuantile()/colorBin()` function to make the colors of the map more intuitive.

3.2.5 Line chart of selected subtheme

In the "Preventieakkoord" of 2018, it is agreed that Dutch people should be healthier by 2040. To reach this goal, measures are being taken in the Netherlands. To see if these measures have any effect, it would be interesting to see the different health variables over time. It would then for example be possible to check whether local campaigns against smoking have an effect on the amount of smokers in an area. So a line chart has been added to facilitate this (box 9 in Figure 5). In the line chart, one line for the selected area is shown, and one for the average values over the years 2012, 2016 and 2020. For the average values, only the areas in the filtered dataset are taken into account. To get this line chart, first the data is filtered such that only the values for the chosen age category remain. Then the average values per year are calculated. Next, the values for the selected area are extracted from the data. These values are all put together in one dataframe which is used as input for the `ggplot()` function. The `fill` argument is used to specify that there should be a differently colored line for the selected area and the average. The function `geom_line` specifies that the plot should be a line chart.

For some variables, there has been very little change over the years. However, in the line chart it sometimes seemed as if there was a very big change but when you look at the y-axis, it could be noticed that there was not even a change of 1%. Users may not look at the axis very closely though, and instead assume that there has been a very big change. Therefore, 5% above and below the y-axis was added to make sure this axis shows at least 10%. This was done by adding the following: `scale_y_continuous(expand = expansion(add = 5))`. Now, the changes seem smaller than before.

3.2.6 Barplot of selected subtheme per age category

To see all age categories in one visualization, the barplot per age category has been included (box 10 in Figure 5). In this plot, the values of the selected subtheme are visible for all age categories for both the selected area and the average. Again, for the average only the areas in the filtered dataset are taken into account. A dataframe with the values for the selected area and for the average has been constructed in a similar way as in the previous section. This dataframe is used as input for the `ggplot()` function. The function `geom_col()` has been used to specify that the heights of the bars have to represent the values in the data.

3.2.7 Barplot per category of selected subtheme

Finally, there is the barplot per category of selected subtheme. This plot only appears when the selected subtheme consists of multiple variables. For example for the subtheme weight that consists of the variables underweight, normal weight, obesity, and morbid obesity, it could be useful to see the values of all the different variables of weight in one plot (see Figure 7). Again, the values are shown for the selected area and the average of the areas in the filtered dataset. For this visualization, a dataframe with the values for the selected area

and the average has been created again. The `geom_col()` function has been used, and in addition the function `coord_flip()` is used to specify that the bars are horizontal instead of vertical.

3.3 General adjustments for increasing user-friendliness

To increase the user-friendliness of the app, some adjustments have been made. These adjustments are applied to both the amenities and the health tab.

For all text that has been used in the app, it was tried to keep it as simple as possible in the hope that it is understandable for most people. All visualizations are also accompanied by some text that explains what can be seen which hopefully results in more clarity as well. Moreover, the size of the labels and legends of the visualizations have been made a bit bigger such that it is readable for everyone.

Furthermore, the loading of the visualizations sometimes take a bit of time. It is possible that users then start to think something is wrong as their desired visualizations do not seem to appear and they can get frustrated. To overcome this problem, a symbol has been added to all plots (using `shinycssloaders::withSpinner()`) that can be seen if the app is still busy loading (see Figure 9). This makes it clear that the app did not get stuck but is simply still loading.

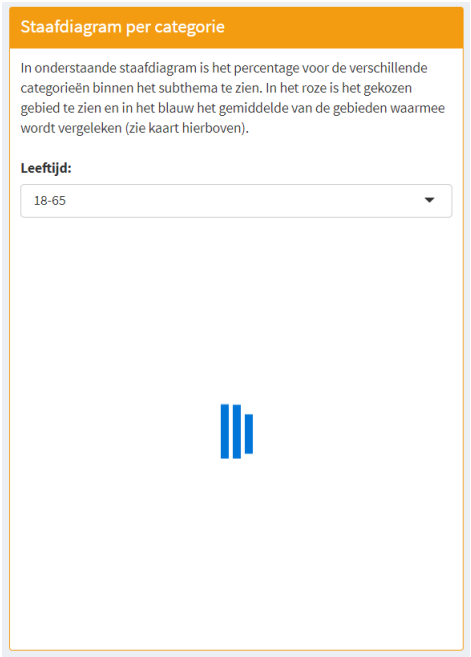


Figure 9: Shown symbol when the visualization is loading.

4 Application Example

In this chapter, some fictional examples are given on how the Shiny app could be used in practice. First, an example for the amenities tab combined with the health tab is provided. Then, an example for only the health tab is given.

4.1 Example 1: amenities and health

A citizen living in Overvecht has noticed that the distance he needs to travel to a hospital seems quite long. He is curious if this is indeed true when Overvecht is compared to other areas. He opens the app and starts filling in the desired input fields. He selects neighbourhood as level, Utrecht as municipality, and Wijk 03 Overvecht as neighbourhood. He wants to compare Overvecht with areas that have the same urbanity level, which is 1. Then he selects *Gezondheid en welzijn* as theme. The subthemes now have both hospitals including external outpatients clinic and hospitals excluding external outpatient clinics as choice. He knows there is an external outpatient clinic close by but this is not what he is searching for and thus selects hospitals excluding external outpatient clinics as the subtheme (*Afstand tot ziekenhuis excl. Buitenpolikliniek (km)*). He then sees in the map with the selected subtheme that the distance people living in Overvecht need to travel to the hospital is

5.5 km, while the average for neighbourhoods with urbanity level 1 is 2.7 km (see Figure 10). His suspicions seem to be proven correct, citizens in Overvecht have to travel a much greater distance than average.

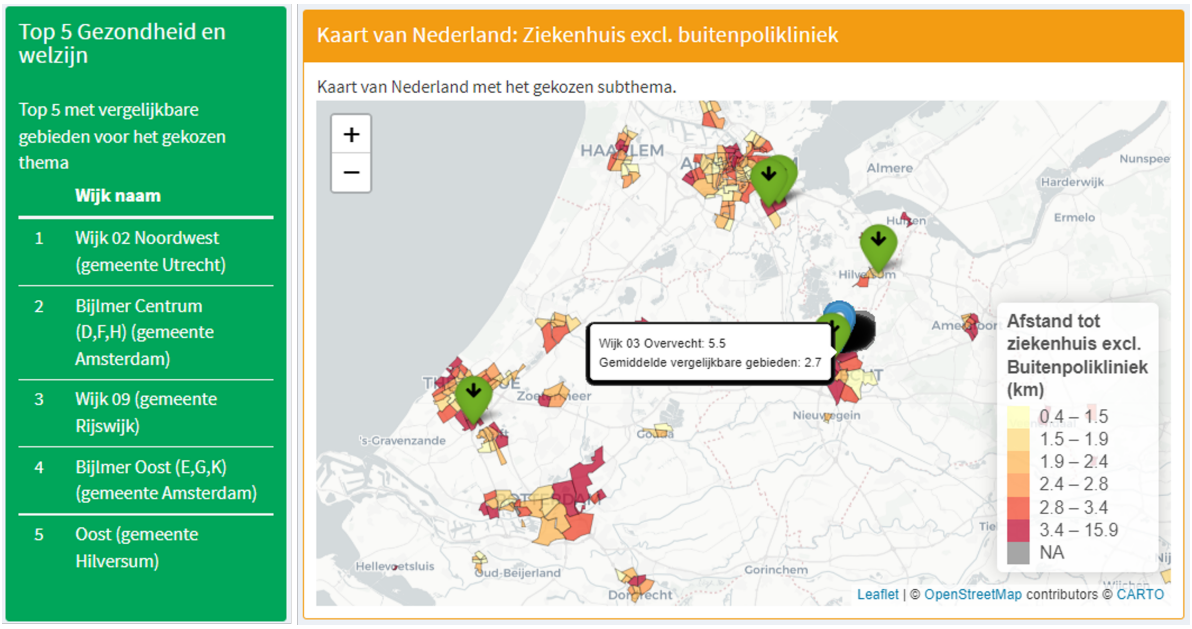


Figure 10: On the right, a map colored by the distance to a hospital (excluding external outpatient clinics) can be seen with the text box pointing to Overvecht. On the left, the 5 most similar areas based on the theme *Gezondheid en welzijn* are shown. These areas are indicated with a green pointer on the map.

He then questions whether this is necessarily a bad thing? Maybe the people living in Overvecht are healthier than average and therefore do not need to go to the hospital as much. To see if this is true, he moves from the amenities tab to the health tab. He again selects neighbourhood Overvecht as desired area. This time, he wants to compare with areas that have a similar age distribution, since health heavily depends on age. As subtheme the percentage of people with very good or good health is selected (*Zeer goede of goede gezondheid (%)*). What he then sees in the histogram (see Figure 11) shocks him a bit.

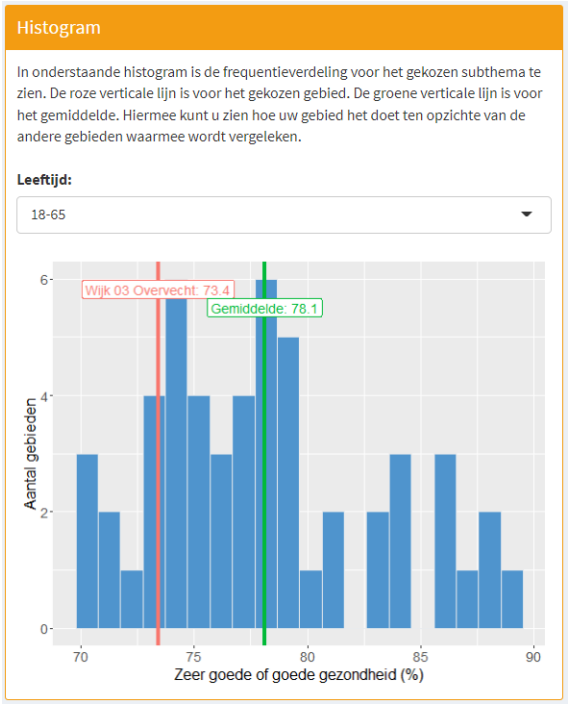


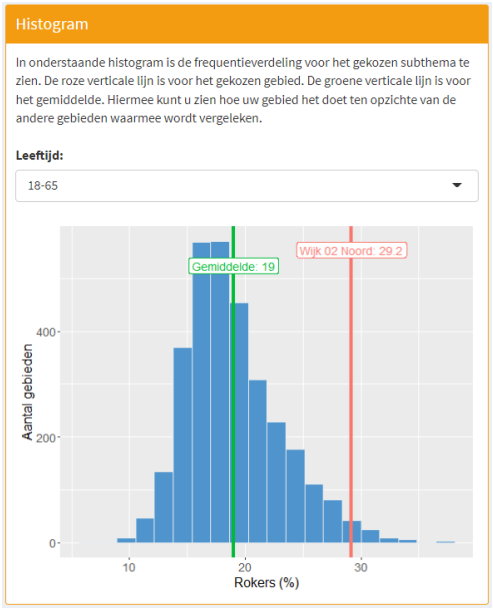
Figure 11: Histogram of people with very good or good health (%) in neighbourhoods with a comparable age distribution as Overvecht. The pink vertical line is for Overvecht, the green line is for the average.

The histogram shows that compared to areas with a similar age distribution, Overvecht has a relatively bad health. In Overvecht, 73.4% of the people have a very good or good health and the average for neighbourhoods with a similar age distribution is 78.1%. So the distance to the hospital is long in Overvecht, while the health of its citizens is relatively low. This does not seem fair, and he decides to bring this up with the municipality to see what their reaction is and whether something can be done about it. He also keeps the five most similar areas based on the theme *Gezondheid en welzijn* in mind (see Figure 10), since this map shows that some of these areas also have a relatively long distance to the hospital. So perhaps he could check how these areas think about this problem and how they deal with it.

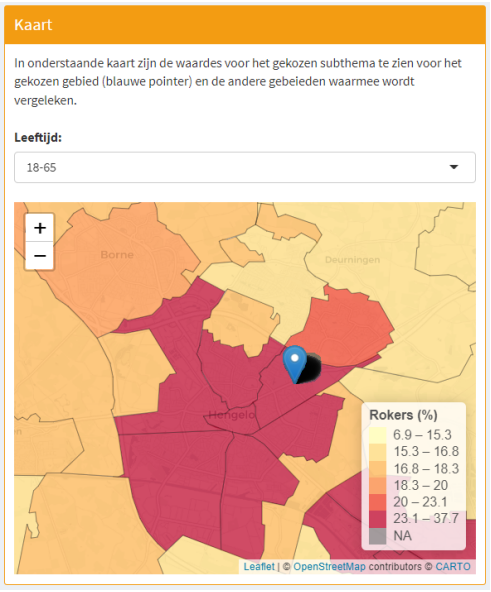
4.2 Example 2: health

In multiple neighbourhoods in municipality Hengelo, a campaign has started against smoking. The municipality claims that the percentage of people smoking in these neighbourhoods is much higher than average. Municipality Hengelo wants to achieve the objective stated in the "Preventieakkoord" that says that less than 5% of the population should smoke by 2040. But to do this, they have some catching up to do.

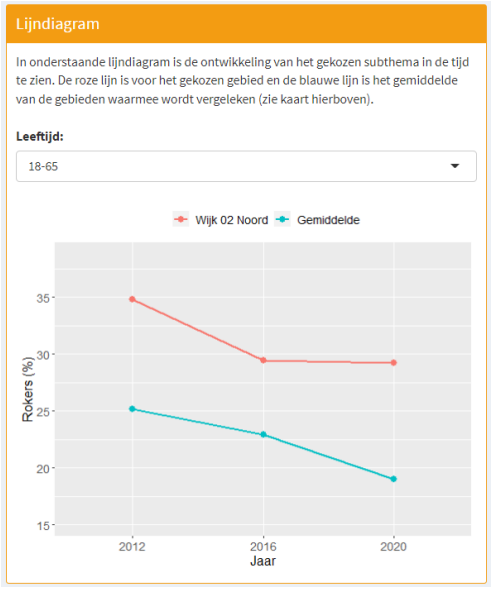
A resident living in one of these neighbourhoods (Wijk 02 Noord) hears about this and is surprised by these claims, she does not think that there is such a high percentage of smokers in her neighbourhood. She is of the opinion that the money spent on this campaign could better be spend on something else. To check the claims of the municipality, she opens the health tab of the app, and compares her neighbourhood with the rest of the Netherlands with the percentage of smokers (*Rokers (%)*) as selected subtheme. The results can be seen in Figure 12.



(a) Histogram of smokers (%) in all neighbourhoods in NL. The pink vertical line is for Wijk 02 Noord and the green line is for the average.



(b) Map of smokers (%) in neighbourhoods in and around Hengelo. The blue pointer shows Wijk 02 Noord.



(c) Line chart of smokers (%) in all neighbourhoods in NL over the years. The red line shows Wijk 02 Noord, blue shows the average.

Figure 12: Results of the health tab for neighbourhood Wijk 02 Noord in municipality of Hengelo compared to all other neighbourhoods in the Netherlands.

The histogram (see Figure 12a) shows her that indeed, they have a very high percentage of smokers compared to other areas in the Netherlands. 29.2% of the population in Wijk 02 Noord are smokers, while the average of all areas in the Netherlands is only 19%. Additionally, the map (see Figure 12b) shows that the municipality was right in their claim that multiple neighbourhoods in Hengelo have a very high percentage of smokers. But what might have surprised her the most in the results, is the line chart (see Figure 12c). In this graph, she sees that over the last four years, the percentage of smokers has decreased on average. Her neighbourhood on the other hand, seems to have no change in the percentage of smokers at all the past four years. However, she knows that the age group 20-40 has the highest percentage of smokers. In the visualization for the age distribution, it can be seen that her neighbourhood has a large percentage of people between 25 and 45 (see Figure 13).

So she decides to compare her neighbourhood with neighbourhoods with a comparable age distribution. The results unfortunately do not show a different outcome. Wijk 02 Noord in Hengelo still has one of the highest

percentage of smokers. Just to be sure, she also compares the neighbourhood with areas that have the same urbanity level (level 2) since she thinks that maybe people living in an urban region smoke more on average. But even then the results do not change. So the municipality was right after all, Wijk 02 Noord in Hengelo has a very high percentage of smokers compared to other areas. Suddenly the campaign does not seem like such a bad idea after all.

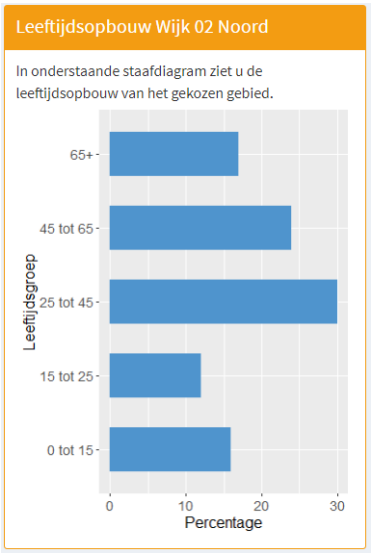


Figure 13: Age distribution of Wijk 02 Noord in municipality Hengelo.

5 Discussion

The application examples demonstrate how the app can be beneficial for people. However, the data and the app also have some limitations. First, the limitations of the used data are discussed, followed by the limitations of the app including some opportunities for further research.

5.1 Data limitations

The first limitation is the missing data for some neighbourhoods and boroughs. Table 2 in the appendix shows that the data about amenities is missing for 0.85% of the neighbourhoods and 4.85% of the boroughs. In table 1a it can be seen that the population is zero in most of these cases. So for these areas the missing data is not a problem. For the other areas, the people who live here cannot use the app for their own area. However, the amount of people for who this is the case is so low that this problem is not that big.

Population	Amount of neighbourhoods	Amount of boroughs
0	21	465
5	6	161
10	-	37
15	-	2
125	-	1
215	-	1
250	-	1
305	-	1
740	-	1

(a) Amenities

Population	Amount of neighbourhoods	Amount of boroughs
0	20	459
5	6	160
10	2	115
15	2	55
20-50	1	28
50-100	-	7
100-200	-	11
200-300	1	8
300-400	-	5
400-500	-	1
500-600	-	2
600-700	-	2
2055	-	1

(b) Health

Table 1: The amount of neighbourhoods and boroughs with a specific population. Contains only neighbourhoods and boroughs that have no data on either amenities (left) or health (right).

For the health data, table 3 shows that the data is missing for 1.48% of the neighbourhoods and 7.92% of the boroughs. In table 1b it can be seen that for most of these areas the population is zero as well and thus not a problem. However, for the health data, there are considerably more areas where the data is missing which have a population of 5 or more. Compared to all citizens in the Netherlands, this is still only a small amount, but for these people the health tab cannot be used for their area.

Additionally, the income level is missing for 8.59% of the neighbourhoods and 31.21% of the boroughs. Especially for the boroughs this is quite a big percentage which is why the option of comparing with areas of the same income level is disabled for boroughs. For neighbourhoods, this option is still possible. So almost 9% of the neighbourhoods has one option less for choosing comparable areas. While this is not ideal, it is assumed that the options of comparing with all other areas in the Netherlands or with the same urbanity level, will be the most used options.

For the health data, the values for the neighbourhoods and boroughs are estimated by the RIVM using a XGBoost model. Therefore, the accuracy of this data can be questioned and conclusions drawn from them might not be entirely right. To keep this uncertainty small, the questionnaires of the RIVM should be completed by as much people as possible. The RIVM has also published data about the 95% confidence intervals of the values. However, RIVM also warns that these intervals are based on assumptions such as the correctness of the model and should thus also be used with caution. It was therefore chosen not to show these values in the visualizations to keep them simple and comprehensible for the average citizen since they might not have enough knowledge to interpret the intervals. To be transparent though, the explanation box clarifies that the data shown in the app are estimates and that their 95% confidence intervals are available online. For people who are interested in these confidence intervals, a visualization could be added to show them. This would however require further research on whether people would be interested in this. And whether the addition of such a visualization would be at the expense of the comprehensibility of the app since the main goal was to create an app that shows the data in an understandable manner. This app could also play a role in checking the accuracy of the XGBoost model. The app could be shown to randomly chosen people and the people could be asked whether they think the data is correct or not. If the answers of people living in areas where the questionnaires have been filled out differ from the answers of people not living in such areas, there is a possibility that the model does not work as expected. In this case, the model should be re-evaluated.

5.2 App limitations and further research opportunities

The app has been shown to a variety of people. Most people found the app comprehensible and were able to understand it, but for some people the app seems to be too difficult. On the "Dag van de Buurt Overvecht", feedback on the app has been collected by conducting informal interviews with interested citizens. These interviews revealed that the app was not that easy to use for some people. Although some adjustments have been done after the interviews, such as adding an explanation with what to do, the app mainly stayed the same. So it might be the case that for people without data literacy skills, the app is too complicated. However, when these people received some extra guidance and were led trough the app by one of the developers, they were able to understand it. Therefore, it might be helpful if data illiterate people can receive help from data intermediaries. If a person has a question that can be answered by the app, they could go to these data intermediaries who can in turn guide them through the app and help explain the results. In this way, the app could also be beneficial for people who otherwise would not be able to understand it.

Another problem of the app is that it can be very slow when the borough level is selected and the user wants to compare with all other boroughs in the Netherlands. Especially the maps take a long time in this case, because 13808 polygons have to be drawn on the map which is a time-consuming process. Sometimes the app cannot even handle this and crashes. So before this app is really used by other people, further research on how to speed it up should be done. One possible option would be to use the function *leafletProxy()* such that the polygons do not have to be reloaded every time a new theme is being chosen. Another option is to load the different visualizations sequentially. While this will not speed up the app, the user experience might be improved by it. Instead of having to wait until all the visualizations are loaded, the user could already look at the ones that have a smaller loading time.

Another point of further research is the addition of more subjects. Right now there are four subjects: amenities, health, traffic accidents (Wooning, 2022), and crime (Kellij, 2022). It would be interesting to have more topics. For example education, housing, or nature. One thing that needs to be kept in mind though, is that all data used for the app cannot contain any personal data. Additionally, the data must be available on borough level. Besides the addition of more subjects, extra years could be included as well. For the amenities, only data about 2020 is taken into account. The extension of earlier years could provide insights on the development over time. The municipal reorganizations over the years do need to be taken into account when data about

the past is used. To keep the app up to date, the addition of later years should be a continuous process. Once data about a new year is available, this should be incorporated in the app. Especially for the health data it would be interesting to see how the values about for example smoking and weight develop over time. Since the "Preventieakkoord" of 2018 states that Dutch people should become more healthy and it would be insightful to discover whether the measures that have been taken to accomplish this, have had any effect.

6 Conclusion

The goal of this research was to visualize open data about amenities and health so that insights can be offered on both the state of the area and the comparison with other areas such that the average citizen is able to make sense of it independently. To reach this goal, a Shiny app has been developed that consists of multiple tabs, including one for amenities and one for health. The presentations of the app to citizens show that for people without any data literacy skills the app might be too difficult. However, there were also people who could make sense of the app independently. So even though the app is not completely comprehensible for all citizens, the average citizen should be able to use the app independently. For those who find the app too difficult, the use of data intermediaries could offer a solution.

The application examples demonstrate how the app could be useful for people. Hopefully, the app can contribute to the improvement of the accessibility of open data, which in turn leads to the possibility of a participatory governance and more transparency.

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory* (pp. 420–434). doi: 10.1007/3-540-44503-X_27

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government information quarterly*, *32*(4), 399–418. doi: 10.1016/j.giq.2015.07.006

CBS. (2020a, September 22). Buurt, wijk en gemeente 2020 voor postcode huisnummer [dataset]. Retrieved 2022-05-01, from https://www.cbs.nl/nl-nl/maatwerk/2020/39/buurt-wijk-en-gemeente-2020-voor-postcode-huisnummer

CBS. (2020b, July 17). Kerncijfers wijken en buurten 2020 [dataset]. Retrieved 2022-04-22, from https://www.cbs.nl/nl-nl/maatwerk/2020/29/kerncijfers-wijken-en-buurten-2020

CBS. (2021, Oct). Toelichting wijk- en buurtkaart, 2019, 2020 en 2021. *Centraal Bureau voor de Statistiek*. Retrieved from https://www.cbs.nl/nl-nl/longread/diversen/2021/toelichting-wijk-en-buurtkaart-2019-2020-en-2021

CBS. (2022). Centraal bureau voor de statistiek. Retrieved 2022-06-10, from https://www.cbs.nl

CBS. (2022, April). Kerncijfers wijken en buurten 2020. Retrieved 2022-04-29, from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/84799NED

Chang, W., & Borges Ribeiro, B. (2021). shinydashboard: Create dashboards with 'shiny'. Retrieved from https://CRAN.R-project.org/package=shinydashboard (R package version 0.7.2)

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., ... Borges, B. (2021). shiny: Web application framework for r. Retrieved from https://CRAN.R-project.org/package=shiny (R package version 1.7.1)

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). doi: 10.1145/2939672.2939785

Cheng, J., Karambelkar, B., & Xie, Y. (2022). leaflet: Create interactive web maps with the javascript 'leaflet' library. Retrieved from https://rstudio.github.io/leaflet/ (R package version 2.1.1)

Curtis, A. E., Smith, T. A., Ziganshin, B. A., & Elefteriades, J. A. (2016). The mystery of the z-score. *Aorta*, *4*(04), 124–130. doi: 10.12945/j.aorta.2016.16.014

Gemeente Utrecht. (2022). Utrecht in cijfers. Retrieved 2022-06-10, from https://utrecht.incijfers.nl

Helden, W. v., Dekker, J., Dorst, P., & van en Govers-Vreeburg, E. (2009). We gooien het de inspraak in. een onderzoek naar de uitgangspunten voor behoorlijke burgerparticipatie. *de Nationale ombudsman*.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, *29*(4), 258–268. doi: 10.1080/10580530.2012.716740

Kellij, S. (2022, July). Compare your neighborhood's amenities and crimes visualizing and comparing open data with r shiny on the municipal, neighborhood and borough level for citizens.

Meijer, A. J., Curtin, D., & Hillebrandt, M. (2012). Open government: connecting vision and voice. *International review of administrative sciences*, *78*(1), 10–29. doi: 10.1177/0020852311429533

OGP. (2022). About open government partnership. Retrieved 2022-06-10, from https://www.opengovpartnership.org/about/

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. Retrieved from https://doi.org/10.32614/RJ-2018-009 (R package version 1.0-7) doi: 10.32614/RJ-2018-009

Rijksoverheid. (2018). Nationaal preventieakkoord: naar een gezonder nederland. Retrieved 2022-05-29, from https://www.rijksoverheid.nl/onderwerpen/gezondheid-en-preventie/documenten/convenanten/2018/11/23/nationaal-preventieakkoord

RIVM. (n.d.). Gezondheid per buurt, wijk en gemeente 2020. Retrieved 2022-06-10, from https://www.rivm.nl/media/smap

RIVM. (n.d.). Verantwoording: gezondheid per buurt, wijk en gemeente. Retrieved 2022-05-31, from https://www.rivm.nl/media/smap/verantwoording.html

RIVM. (2021). Rijksinstituut voor volksgezondheid en milieu. Retrieved 2022-06-10, from https://www.rivm.nl

RIVM. (2022, February 22). Gezondheid per wijk en buurt; 2012/2016/2020 (indeling 2020) [dataset]. Retrieved from https://statline.rivm.nl/portal.html?_la=nl&_catalog=RIVM&tableId=50090NED&_theme=85

Ruijer, E., & Dymanus, C. (2022, May 09). Input voor de visie. , 6-7. (Samen voor Overvecht)

Ruijer, E., Grimmelikhuijsen, S., & Meijer, A. (2017). Open data for democracy: Developing a theoretical framework for open data use. *Government Information Quarterly*, *34*(1), 45–52. doi: 10.1016/j.giq.2017.01.001

Safarov, I., Meijer, A., & Grimmelikhuijsen, S. (2017). Utilization of open government data: A systematic literature review of types, conditions, effects and users. *Information Polity*, *22*(1), 1–24. doi: 10.3233/IP-160012

Samen voor Overvecht. (2022). Voortgangsrapportage 2022 samen voor overvecht.

Teucher, A., & Russell, K. (2021). rmapshaper: Client for 'mapshaper' for 'geospatial' operations. Retrieved from https://CRAN.R-project.org/package=rmapshaper (R package version 0.4.5)

Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). Benchmarking open government: An open data perspective. *Government Information Quarterly*, *31*(2), 278–290. doi: 10.1016/j.giq.2013.10.011

VNG. (2022). Waarstaatjegemeente.nl. Retrieved 2022-06-10, from https://www.waarstaatjegemeente.nl

Wegdam, E. (2022a, April). Kort verslag inspiratiesessie 4 einsteinkwartier/centrum overvecht. (Samen voor Overvecht)

Wegdam, E. (2022b, March). Verslag inspiratiesessie 3 einsteinkwartier. (Samen voor Overvecht)

Wesselink, F. (2020, May). Lokaal/regionaal preventieakkoord: Heeft het meerwaarde? *GGD GHOR*.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Retrieved from https://ggplot2.tidyverse.org (R package version 3.3.5)

Wickham, H. (2021). *Mastering shiny.* O'Reilly Media, Inc.

Wooning, P. (2022, July). Comparisons on municipality, neighborhood and borough level made possible. an r shiny tool to empower citizens with open data about amenities and traffic incidents.

Yoon, A., & Copeland, A. (2019). Understanding social impact of data on local communities. *Aslib Journal of Information Management*, *71*(4), 558–567. doi: 10.1108/AJIM-12-2018-0310

A Used variables and missing data

The following table shows all variables from the amenities dataset that were used for this research. Most of the variables were renamed such that they have clear and informative names. The variables include the names and codes of the areas, the urbanity level, the age distribution, households under or around the social minimum, and the amenities variables. Additionally, the amount of missing data for these variables is included. These were calculated separately for municipalities, neighbourhoods and boroughs.

Variable	Missing for municipality (%)	Missing for neighbourhood (%)	Missing for borough (%)
Gemeentenaam	0.00	0.00	0.00
Gemeentecode	0.00	0.00	0.00
Wijknaam	-	0.00	0.00
Wijkcode	-	0.00	0.00
Buurtnaam	-	-	0.00
Buurtcode	-	-	0.00
Stedelijkheid (1=zeer sterk stedelijk, 5=niet stedelijk)	0.00	0.09	0.50
Personen 0 tot 15 jaar (%)	0.00	2.23	10.57
Personen 15 tot 25 jaar (%)	0.00	2.23	10.57
Personen 25 tot 45 jaar (%)	0.00	2.23	10.57
Personen 45 tot 65 jaar (%)	0.00	2.23	10.57
Personen 65 jaar en ouder (%)	0.00	2.23	10.57
Huishoudens onder of rond sociaal minimum (%)	0.00	8.59	31.21
Afstand tot huisartsenpraktijk (km)	0.00	0.85	4.85
Aantal huisartsenpraktijken binnen 1 km	0.00	0.85	4.85
Aantal huisartsenpraktijken binnen 3 km	0.00	0.85	4.85
Aantal huisartsenpraktijken binnen 5 km	0.00	0.85	4.85
Afstand tot huisartsenpost (km)	0.00	0.85	4.85
Afstand tot apotheek (km)	0.00	0.85	4.85
Afstand tot ziekenhuis incl. buitenpolikliniek (km)	0.00	0.85	4.85
Aantal ziekenhuizen incl. buitenpolikliniek binnen 5 km	0.00	0.85	4.85
Aantal ziekenhuizen incl. buitenpolikliniek binnen 10 km	0.00	0.85	4.85
Aantal ziekenhuizen incl. buitenpolikliniek binnen 20 km	0.00	0.85	4.85
Afstand tot ziekenhuis excl. Buitenpolikliniek (km)	0.00	0.85	4.85
Aantal ziekenhuizen excl. Buitenpolikliniek binnen 5 km	0.00	0.85	4.85
Aantal ziekenhuizen excl. Buitenpolikliniek binnen 10 km	0.00	0.85	4.85
Aantal ziekenhuizen excl. Buitenpolikliniek binnen 20 km	0.00	0.85	4.85
Afstand tot grote supermarkt (km)	0.00	0.85	4.85
Aantal grote supermarkten binnen 1 km	0.00	0.85	4.85
Aantal grote supermarkten binnen 3 km	0.00	0.85	4.85
Aantal grote supermarkten binnen 5 km	0.00	0.85	4.85
Afstand tot overige dagelijkse levensmiddelen (km)	0.00	0.85	4.85
Aantal winkels overige dagelijkse levensmiddelen binnen 1 km	0.00	0.85	4.85
Aantal winkels overige dagelijkse levensmiddelen binnen 3 km	0.00	0.85	4.85
Aantal winkels overige dagelijkse levensmiddelen binnen 5 km	0.00	0.85	4.85
Afstand tot warenhuis (km)	0.00	0.85	4.85
Aantal warenhuizen binnen 5 km	0.00	0.85	4.85
Aantal warenhuizen binnen 10 km	0.00	0.85	4.85
Aantal warenhuizen binnen 20 km	0.00	0.85	4.85
Afstand tot cafetaria (km)	0.00	0.85	4.85
Aantal cafetaria's binnen 1 km	0.00	0.85	4.85
Aantal cafetaria's binnen 3 km	0.00	0.85	4.85
Aantal cafetaria's binnen 5 km	0.00	0.85	4.85
Afstand tot restaurant (km)	0.00	0.85	4.85
Aantal restaurants binnen 1 km	0.00	0.85	4.85
Aantal restaurants binnen 3 km	0.00	0.85	4.85
Aantal restaurants binnen 5 km	0.00	0.85	4.85
Afstand tot hotel (km)	0.00	0.85	4.85

Aantal hotel binnen 5 km	0.00	0.85	4.85
Aantal hotel binnen 10 km	0.00	0.85	4.85
Aantal hotel binnen 20 km	0.00	0.85	4.85
Afstand tot kinderdagverblijf (km)	0.00	0.85	4.85
Aantal kinderdagverblijf binnen 1 km	0.00	0.85	4.85
Aantal kinderdagverblijf binnen 3 km	0.00	0.85	4.85
Aantal kinderdagverblijf binnen 5 km	0.00	0.85	4.85
Afstand tot buitenschoolse opvang (km)	0.00	0.85	4.85
Aantal buitenschoolse opvang binnen 1 km	0.00	0.85	4.85
Aantal buitenschoolse opvang binnen 3 km	0.00	0.85	4.85
Aantal buitenschoolse opvang binnen 5 km	0.00	0.85	4.85
Afstand tot basisscholen (km)	0.00	0.85	4.85
Aantal basisscholen binnen 1 km	0.00	0.85	4.85
Aantal basisscholen binnen 3 km	0.00	0.85	4.85
Aantal basisscholen binnen 5 km	0.00	0.85	4.85
Afstand tot voortgezet onderwijs (km)	0.00	0.85	4.85
Aantal voortgezet onderwijs binnen 3 km	0.00	0.85	4.85
Aantal voortgezet onderwijs binnen 5 km	0.00	0.85	4.85
Aantal voortgezet onderwijs binnen 10 km	0.00	0.85	4.85
Afstand tot scholen VMBO (km)	0.00	0.85	4.85
Aantal scholen VMBO binnen 3 km	0.00	0.85	4.85
Aantal scholen VMBO binnen 5 km	0.00	0.85	4.85
Aantal scholen VMBO binnen 10 km	0.00	0.85	4.85
Afstand tot scholen HAVO/VWO (km)	0.00	0.85	4.85
Aantal scholen HAVO/VWO binnen 3 km	0.00	0.85	4.85
Aantal scholen HAVO/VWO binnen 5 km	0.00	0.85	4.85
Aantal scholen HAVO/VWO binnen 10 km	0.00	0.85	4.85
Afstand tot brandweerkazerne (km)	0.00	0.85	4.85
Afstand tot oprit hoofdverkeersweg (km)	0.00	0.85	4.85
Afstand tot treinstation (km)	0.00	0.85	4.85
Afstand tot belangrijk overstapstation (km)	0.00	0.85	4.85
Afstand tot zwembad (km)	0.00	0.85	4.85
Afstand tot kunstijsbaan (km)	0.00	0.85	4.85
Afstand tot bibliotheek (km)	0.00	0.85	4.85
Afstand tot poppodium (km)	0.00	0.85	4.85
Afstand tot bioscoop (km)	0.00	0.85	4.85
Aantal bioscoop binnen 5 km	0.00	0.85	4.85
Aantal bioscoop binnen 10 km	0.00	0.85	4.85
Aantal bioscoop binnen 20 km	0.00	0.85	4.85
Afstand tot sauna (km)	0.00	0.85	4.85
Afstand tot zonnebank (km)	0.00	0.85	4.85
Afstand tot attractie (km)	0.00	0.85	4.85
Aantal attracties binnen 10 km	0.00	0.85	4.85
Aantal attracties binnen 20 km	0.00	0.85	4.85
Aantal attracties binnen 50 km	0.00	0.85	4.85
Afstand tot podiumkunsten (km)	0.00	0.85	4.85
Aantal podiumkunsten binnen 5 km	0.00	0.85	4.85
Aantal podiumkunsten binnen 10 km	0.00	0.85	4.85
Aantal podiumkunsten binnen 20 km	0.00	0.85	4.85
Afstand tot museum (km)	0.00	0.85	4.85
Aantal musea binnen 5 km	0.00	0.85	4.85
Aantal musea binnen 10 km	0.00	0.85	4.85
Aantal musea binnen 20 km	0.00	0.85	4.85

Table 2: All used variables and the percentage they are missing from the amenities dataset for the municipalities, neighbourhoods and boroughs.

The following three tables show the used variables from the health dataset, together with the amount of missing data for the years 2020, 2016, and 2012. These were again calculated separately for municipalities, neighbourhoods and boroughs. For the years 2016 and 2012, some variables have a missing percentage of 100. This means that the questions from the questionnaire of these years were either not asked or not comparable with the questions from 2020.

Table for 2020:

Variable	**Missing for municipality (%)**	**Missing for neighbourhood (%)**	**Missing for borough (%)**
Zeer goede of goede gezondheid (%)	0.00	1.48	7.92
Voldoet aan beweegrichtlijn (%)	0.00	1.48	7.92
Wekelijkse sporters (%)	0.00	1.48	7.92
Ondergewicht (%)	0.00	1.48	7.92
Normaal gewicht (%)	0.00	1.48	7.92
vergewicht (%)	0.00	1.48	7.92
Ernstig overgewicht (%)	0.00	1.48	7.92
Rokers (%)	0.00	1.48	7.92
Voldoet aan alcoholrichtlijn (%)	0.00	1.48	7.92
Drinkers (%)	0.00	1.48	7.92
Zware drinkers (%)	0.00	1.48	7.92
Overmatige drinkers (%)	0.00	1.48	7.92
Langdurige aandoening (%)	0.00	1.48	7.92
Beperkt vanwege gezondheid (%)	0.00	1.48	7.92
Ernstig beperkt vanwege gezondheid (%)	0.00	1.48	7.92
Langdurige ziekte en beperkt (%)	0.00	1.48	7.92
Lichamelijke beperking (%)	0.00	1.48	7.92
Beperking in horen (%)	0.00	1.48	7.92
Beperking in zien (%)	0.00	1.48	7.92
Beperking in bewegen (%)	0.00	1.48	7.92
Matig tot hoog risico op angststoornis of depressie (%)	0.00	1.48	7.92
Hoog risico op angststoornis of depressie (%)	0.00	1.48	7.92
(heel) veel stress (%)	0.00	1.48	7.92
Matig tot veel regie over eigen leven (%)	0.00	1.48	7.92
Eenzaam (%)	0.00	1.48	7.92
Ernstig eenzaam (%)	0.00	1.48	7.92
Emotioneel eenzaam (%)	0.00	1.48	7.92
Sociaal eenzaam (%)	0.00	1.48	7.92
Mantelzorger (%)	0.00	1.48	7.92
Vrijwilligerswerk (%)	0.00	1.48	7.92
Lopen en of fietsen naar school of werk (%) *	66.67	67.00	68.73
Lopen naar school of werk (%) *	66.67	67.00	68.73
Fietsen naar school of werk (%) *	66.67	67.00	68.73
Ernstige geluidhinder door buren (%)	0.00	1.48	7.92
Moeite met rondkomen (%)	0.00	1.48	7.92

Table 3: Percentages of the missing data from the health dataset of 2020 per used variable for the municipalities, neighbourhoods and boroughs.
* The variables about biking/cycling to school/work are only available for age class 18-65, hence the high percentages of missing data for these variables

Table for 2016:

Variable	**Missing for municipality (%)**	**Missing for neighbourhood (%)**	**Missing for borough (%)**
Zeer goede of goede gezondheid (%)	0.00	1.62	8.56
Voldoet aan beweegrichtlijn (%)	0.00	1.62	8.56
Wekelijkse sporters (%)	0.00	1.62	8.56
Ondergewicht (%)	0.00	1.62	8.56
Normaal gewicht (%)	0.00	1.62	8.56
Overgewicht (%)	0.00	1.62	8.56
Ernstig overgewicht (%)	0.00	1.62	8.56
Rokers (%)	0.00	1.62	8.56
Voldoet aan alcoholrichtlijn (%)	0.00	1.62	8.56
Drinkers (%)	0.00	1.62	8.56
Zware drinkers (%)	0.00	1.62	8.56
Overmatige drinkers (%)	0.00	1.62	8.56
Langdurige aandoening (%)	0.00	1.62	8.56
Beperkt vanwege gezondheid (%)	100.00	100.00	100.00
Ernstig beperkt vanwege gezondheid (%)	100.00	100.00	100.00
Langdurige ziekte en beperkt (%)	100.00	100.00	100.00
Lichamelijke beperking (%)	0.00	1.62	8.56
Beperking in horen (%)	0.00	1.62	8.56
Beperking in zien (%)	0.00	1.62	8.56
Beperking in bewegen (%)	0.00	1.62	8.56
Matig tot hoog risico op angststoornis of depressie (%)	0.00	1.62	8.56
Hoog risico op angststoornis of depressie (%)	0.00	1.62	8.56
(heel) veel stress (%)	100.00	100.00	100.00
Matig tot veel regie over eigen leven (%)	0.00	1.62	8.56
Eenzaam (%)	0.00	1.62	8.56
Ernstig eenzaam (%)	0.00	1.62	8.56
Emotioneel eenzaam (%)	0.00	1.62	8.56
Sociaal eenzaam (%)	0.00	1.62	8.56
Mantelzorger (%)	0.00	1.62	8.56
Vrijwilligerswerk (%)	0.00	1.62	8.56
Lopen en of fietsen naar school of werk (%) *	66.67	67.01	68.80
Lopen naar school of werk (%) *	66.67	67.01	68.80
Fietsen naar school of werk (%) *	66.67	67.01	68.80
Ernstige geluidhinder door buren (%)	100.00	100.00	100.00
Moeite met rondkomen (%)	0.00	1.62	8.56

Table 4: Percentages of the missing data from the health dataset of 2016 per used variable for the municipalities, neighbourhoods and boroughs.
* The variables about biking/cycling to school/work are only available for age class 18-65, hence the high percentages of missing data for these variables

Table for 2012:

Variable	**Missing for municipality (%)**	**Missing for neighbourhood (%)**	**Missing for borough (%)**
Zeer goede of goede gezondheid (%)	0.00	1.77	9.28
Voldoet aan beweegrichtlijn (%)	100.00	100.00	100.00
Wekelijkse sporters (%)	100.00	100.00	100.00
Ondergewicht (%)	0.00	1.77	9.28
Normaal gewicht (%)	0.00	1.77	9.28
Overgewicht (%)	0.00	1.77	9.28
Ernstig overgewicht (%)	0.00	1.77	9.28
Rokers (%)	0.00	1.77	9.28
Voldoet aan alcoholrichtlijn (%)	100.00	100.00	100.00
Drinkers (%)	0.00	1.77	9.28
Zware drinkers (%)	0.00	1.77	9.28
Overmatige drinkers (%)	0.00	1.77	9.28
Langdurige aandoening (%)	100.00	100.00	100.00
Beperkt vanwege gezondheid (%)	100.00	100.00	100.00
Ernstig beperkt vanwege gezondheid (%)	100.00	100.00	100.00
Langdurige ziekte en beperkt (%)	100.00	100.00	100.00
Lichamelijke beperking (%)	0.00	1.77	9.28
Beperking in horen (%)	0.00	1.77	9.28
Beperking in zien (%)	0.00	1.77	9.28
Beperking in bewegen (%)	0.00	1.77	9.28
Matig tot hoog risico op angststoornis of depressie (%)	0.00	1.77	9.28
Hoog risico op angststoornis of depressie (%)	0.00	1.77	9.28
(heel) veel stress (%)	100.00	100.00	100.00
Matig tot veel regie over eigen leven (%)	100.00	100.00	100.00
Eenzaam (%)	0.00	1.77	9.28
Ernstig eenzaam (%)	0.00	1.77	9.28
Emotioneel eenzaam (%)	0.00	1.77	9.28
Sociaal eenzaam (%)	100.00	100.00	100.00
Mantelzorger (%)	0.00	1.77	9.28
Vrijwilligerswerk (%)	100.00	100.00	100.00
Lopen en of fietsen naar school of werk (%) *	66.67	66.98	68.88
Lopen naar school of werk (%) *	66.67	66.98	68.88
Fietsen naar school of werk (%) *	66.67	66.98	68.88
Ernstige geluidhinder door buren (%)	100.00	100.00	100.00
Moeite met rondkomen (%)	0.00	1.77	9.28

Table 5: Percentages of the missing data from the health dataset of 2012 per used variable for the municipalities, neighbourhoods and boroughs.
* The variables about biking/cycling to school/work are only available for age class 18-65, hence the high percentages of missing data for these variables

Description of all variables in the health dataset:

Variable name	Variable description
Zeer goede of goede gezondheid (%)	People who consider own health as very good or good (%)
Voldoet aan beweegrichtlijn (%)	People who perform at least 150 minutes per week of moderately intensive exercise (%)
Wekelijkse sporters (%)	People who exercise at least once a week (%)
Ondergewicht (%)	People who are underweight (BMI lower than 18.5) (%)
Normaal gewicht (%)	People who have a healthy weight (BMI between 18.5 and 25) (%)
Overgewicht (%)	People who are overweight (BMI 25 or higher) (%)
Ernstig overgewicht (%)	People who are extremely overweight (BMI 30 or higher) (%)
Rokers (%)	People who smoke (%)
Voldoet aan alcoholrichtlijn (%)	People who comply with the alcohol directive (max 1 glass a day) (%)
Drinkers (%)	People who sometimes drink alcohol (%)
Zware drinkers (%)	People who are heavy drinkers (drinking at least 1 times a week at least 6 (M) or 4 (F) glasses per day) (%)
Overmatige drinkers (%)	People who are excessive drinkers (more than 21 (M) or 14 (F) glasses per week) (%)
Langdurige aandoening (%)	People who have one or more longterm illness (6 months or longer) (%)
Beperkt vanwege gezondheid (%)	People who are limited in daily life because of health problems (%)
Ernstig beperkt vanwege gezondheid (%)	People who are seriously limited in daily life because of health problems (%)
Langdurige ziekte en beperkt (%)	People who are limited in daily life because of problems with health 6 months or longer (%)
Lichamelijke beperking (%)	People with a hearing, vision, or mobility disability (%)
Beperking in horen (%)	People with a hearing disability (%)
Beperking in zien (%)	People with a vision disability (%)
Beperking in bewegen (%)	People with a mobility disability (%)
Matig tot hoog risico op angststoornis of depressie (%)	People with a moderate to high risk of anxiety disorder or depression (%)
Hoog risico op angststoornis of depressie (%)	People with a high risk of anxiety disorder or depression (%)
(heel) veel stress (%)	People who experienced a lot of stress in the past 4 weeks (%)
Matig tot veel regie over eigen leven (%)	People who feel a moderate to high control over their own life (%)
Eenzaam (%)	People who are lonely (%)
Ernstig eenzaam (%)	People who are seriously lonely (%)
Emotioneel eenzaam (%)	People who are emotionally lonely (%)
Sociaal eenzaam (%)	People who are socially lonely (%)
Mantelzorger (%)	People who are caregivers (at least 3 months and/or at least 8 hours a week) (%)
Vrijwilligerswerk (%)	People who do volunteer work (%)
Lopen en of fietsen naar school of werk (%)	People who (partly) bike or walk to school or work at least 1 day a week (%)
Lopen naar school of werk (%)	People who (partly) walk to school or work at least 1 day a week (%)
Fietsen naar school of werk (%)	People who (partly) bike to school or work at least 1 day a week (%)
Ernstige geluidhinder door buren (%)	People who experience severe noise disturbance by neighbours (%)
Moeite met rondkomen (%)	People who experience difficulty with household income (%)

Table 6: Variable names and their description of the health dataset.

B Open code

All code used for preprocessing the data and developing the app is openly available and can be found on Github: https://github.com/ImkeDekkers/Buurtvergelijker

The app itself can be found in the file *Shiny_ app*. The code specifically for the amenities tab in the ui and server starts with "### Facilities". Separate files such as for the data preparation and visualizations for amenities can be found in the folder *Facilities*. The code specifically for the health tab in the ui and server starts with "### Health". The separate files for the data preparation and visualizations can be found in the *Health* folder.