MASTER THESIS

Comparisons on municipality, neighborhood and borough level made possible

An R Shiny tool to empower citizens with open data about amenities and traffic incidents

8982643 | Pascalle Kelsey Wooning

Master Applied Data Science (MSc)
Utrecht University
Faculty of Science
Graduate School of Natural Sciences

July 1st 2022

Supervisors:
L. Boeschoten PhD
E. van Kesteren PhD

Abstract

Active participation of citizens is more and more stimulated by governmental institutions. For example, citizens are allowed to contribute to the decision-making process by submitting ideas on how to improve the livability of the neighborhood. Most of the community initiatives are experience-based and not evidence-based or data driven. However, it is not easy for citizens to obtain objective data about different topics of interest. Multiple online dashboards fail to capture and visualize data of boroughs and comparisons between different geographical areas are not possible. Therefore, the current study tries to bridge this gap by developing a dashboard with visualizations of open data and comparisons between geographical areas about topics of interest for citizens. The study focuses on open data about amenities and traffic incidents as was suggested by citizens of Utrecht Overvecht. Direct feedback from citizens of Utrecht Overvecht and scientific researchers contributed to the development process of the dashboard. The feedback resulted in improved understandability and usability of the dashboard designed for citizens of the Netherlands. The written code to develop the dashboard is published on GitHub to contribute to open science.

Keywords: R Shiny, community initiatives, open data, visualizations

Table of contents

1. Introduction

The question of what the link is between the wisdom of the crowd and citizen participation, can be answered with two words: community initiatives. A community initiative is any voluntary action or idea by citizens that are more or less directly aimed at influencing the management of collective affairs or improving the living environment of the neighborhood (Verba, Schlozman & Brady, 1995; Movisie, 2012). There are several benefits of citizen participation for both citizens and the municipality. For example, when citizens actively participate in the democratic processes, it allows for inclusion and voices to be heard and can result in feelings of responsibility (Michels & De Graaf, 2010). The municipality can benefit from the wisdom of the crowd by allowing citizens to share their perspective of and potential solutions to neighborhood needs (Michels & De Graaf, 2010; Ruiter, Grimmelinkhuijsen & Meijer, 2017; Wilson, 2019). Besides, the municipality benefits from citizen participation because it contributes to a greater legitimacy of policy decisions (Michels & De Graaf, 2010; Kim & Lee, 2019).

Municipalities recognize the importance of community initiatives as the call for community initiatives is advertised on websites of Dutch municipalities. On the websites it is stated clearly that good community initiatives can be funded by the government. In total, Amsterdam budgeted 1,2 million euros for community initiatives to improve the living environment (Heida, 2022). In Oosterhout, 300.000 euros are set aside to enrich neighborhoods with more greenery (Oosterhout Nieuws, 2021). Other examples of community initiatives are the construction of a Jeu de Boules alley in Arnhem, an environmental roundabout or allotment garden in Soest and Steenwijk (Arnhemse Koerier, 2022; Bolt, 2022; Smit, 2022). These initiatives grow from the needs of inhabitants as they experience life in the neighborhood. However, when presented to the municipality, neighborhood initiatives may be turned down, because the plan does not meet the requirements of the municipality. Therefore, it is important to hand citizens the tools to reinforce their standing point.

Currently, citizens often initiate projects based on their own experiences (Yoon & Copeland, 2019). Citizens who are actively engaging in neighborhood activities have a great understanding of what the neighborhood needs to improve the living environment, but ideas for improvement are mostly based on feelings. Furthermore, it may be that experience from inhabitants is colored by their own perceptions and beliefs and therefore inaccurately represents the greater good of the whole community (Yoon & Copeland, 2019). Therefore, experience-based knowledge should be complemented with data to objectify the issues and needs of the neighborhood area. The use of data by citizens can drive the change from an experience-based approach to a more data-driven and evidence-based approach (Luthfi & Janssen, 2019). Governmental institutions use data to gain insights in what the neighborhood needs and to justify certain decisions about the neighborhood. These data can be in contrast with the experiences of citizens. Therefore, being open about the used data as well as disclosing the data entirely can contribute to the understanding of citizens about governmental decisions (Wessels, Finn, Sveinsdottir & Wadhwa, 2017; Luthfi & Janssen, 2019).

When the data is available to citizens, they can actively participate in the decision-making process in a meaningful way (Meijer, Curtin & Hillebrandt, 2012; Wessels et al., 2017). Citizens can voice their opinions about what the data are telling about the neighborhood and complement this knowledge with their experiences in the neighborhood. Furthermore, citizens can check the government and understand why certain decisions are taken when decisions do not meet their expectations (Meijer et al., 2012; Wessels et al., 2017). The use and publication of open data contributes to the quality of governance and public trust in governmental decisions (Ruiter et al., 2017; Wilson, 2019). Open data refers to data that are openly available, accessible, understandable and

reusable (Wessels et al., 2017). Access to open data is an important aspect of democracy and the government has the responsibility to disclose information (Ruiter et al., 2017). In recent years, the governmental institutions recognized the importance of open data, which resulted in more and more governments that support inclusive decision-making by publishing datasets (Ubaldi, 2019).

If data are accessible to citizens, it is important that citizens understand the data. However, understanding data can be difficult for citizens as a certain amount of knowledge and skills are required to correctly interpret the data (Montes & Slater, 2019; Yoon & Copeland, 2019). Only with a certain level of data literacy, citizens can translate numbers and figures into specific demands or public interests (Ruiter et al., 2017). Available online dashboards such as allecijfers.nl or incijfers.nl try to help citizens better understand and interpret data by visualizing the data. However, the already published online dashboards do not meet the needs of citizens. Therefore, the current study tries to connect citizen needs to open data for livability improvements in the neighborhood.

One way in which the already published online dashboard do not meet the needs of citizens, is that data and visualizations are only accessible for a small number of municipalities. The unavailability of data and visualizations makes it impossible for citizens to objectively obtain knowledge about the neighborhood based on open data. Furthermore, the existing dashboards only visualize data about the selected municipality. As comparisons with other municipalities are impossible on the existing dashboards, citizens have no clue if particular values and percentages of variables are high or low compared to similar municipalities. Also, the visualizations can only be presented on the level of the municipality or neighborhood, whereas insights into boroughs can be valuable for specific community initiatives as each borough has other needs.

The current study tries to connect citizen needs for knowledge and livability improvements in the neighborhood to open data as part of the community initiative *"Samen voor Overvecht*" (Samen voor Overvecht, 2022). Citizens have voiced the needs of Overvecht in a meeting with Wijkplatform Overvecht, Bewonersplatform Overvecht and the municipality of Utrecht (Wegdam, 2022). Among other topics, citizens were interested in what amenities are present in the neighborhood as well as the distance to these amenities. When the data can indicate a lack of certain amenities in the neighborhood, this can strengthen the voice of citizens in their debate with the municipality about their needs. Furthermore, citizens of Utrecht Overvecht are wondering how the road safety of Overvecht compares to other, similar neighborhoods (Wegdam, 2022). Insights in traffic incidents can drive decisions about road obstacles or adjustments of speed limit in the area. The need for safer roads in Overvecht can also be concluded from adjustments in different traffic situations. The speed limit of multiple streets has been adjusted from 50 kilometers per hour to 30 kilometers per hour (Echt Overvecht, 2022). The way in which the road was made safer was introduced by a company in Overvecht, which won a contest on the topic of road safety, and is an example of citizen participation.

Both questions about amenities and road safety could not be answered using already existing online dashboards. Therefore, a dashboard is developed which accounts for the shortcomings of existing dashboards and closes the gap between the needs of citizens and open data visualizations. As the dashboard is developed to meet the needs of citizens of Utrecht Overvecht specifically, and citizens of the Netherlands in general, it is important that the right visualizations are chosen. The visualizations serve to indicate needs and issues in the neighborhood and to compare geographical areas to similar areas across the Netherlands. Therefore, the current study focused on the following research question: '*How can open data be visualized as accurately and comprehensible as possible for citizens of the Netherlands*?'.

To answer the research question, an iterative study is conducted. The process resulted in going back and forth in deciding what is the best approach to make data accessible and understandable for citizens. Citizens of Utrecht Overvecht are actively involved in the research through providing feedback on the developed dashboard. The dashboard is constructed using a tool named R shiny (Chang et al., 2021), which makes it possible to create an interactive environment where data can be filtered and displayed as desired. R shiny is a simple tool with lots of possibilities for customization to meet the research goals and needs of the citizens. The written code is published on GitHub[1] as contribution to open science.

In the next section it is discussed what data are used and which approach is taken to prepare the data for visualization. This data section will include what considerations are taken into account when preprocessing the data into a proper format. The third section contains information on the construction of the dashboard and how the chosen variables can be best visualized. Then, an example case study is given to show how the dashboard can be used. The paper will conclude with a discussion about the taken decisions, limitations and suggestions for further research.

[1] https://github.com/ImkeDekkers/Buurtvergelijker. An explanation of the structure of this repository can be found in Appendix V

2. Open data

The current section explains the open data are used. Furthermore, it discusses important properties of the data, such as descriptions of the included variables and how the data are gathered and preprocessed. Section 2.1. focuses on the open data of Statistics Netherlands (CBS) about amenities. Section 2.2. discusses the open data of the Department of Waterways and Public Works (RWS) about traffic incidents. Both datasets contain more variables than needed for the current study. Therefore only included variables are covered in the next sections.

2.1. Amenities; using the CBS-dataset

2.1.1. Variables of interest

As citizens are interested in the proximity of amenities, the distance to these amenities need to be included in the dataset. CBS (CBS, 2021a) provides information about the proximity of amenities in open data. The distance to amenities is calculated for each inhabitant of a specified area over paved car roads. Then, the average is calculated of all the distances from people's addresses in a particular area to the amenities (CBS, n.d.). This average is registered in the data as a decimal number for almost all geographical areas. Besides the *average distance* to amenities, the average *number of amenities* within a fixed range is calculated and provided by CBS. The average *number of amenities* is calculated per person and averaged over the number of inhabitants in a geographical area (CBS, n.d.). The fixed distance can either be 1, 3, 5, 10, 20 or 50 kilometer. The choice for a certain range is dependent on the density of amenities and specified in Appendix I. In short, the variables of interest include the *average distance* to amenities and the *number of amenities* within a specified range.

The *average distance* to and *number of amenities* give insights into the approximate effort to reach a certain facility. In an ideal world, the distance to all amenities is short and makes sure that the amenities are reached easily from each corner of the area. Including the *average distance* to and *number of amenities* in the dashboard indicates if particular amenities need to be relocated or initiated in a particular area. Knowledge about this information can be helpful in initiating community initiatives.

As mentioned before, the study aims to close the gap between the needs of citizens and the visualized open data in online dashboards. In doing so, the developed dashboard makes comparisons between geographical areas and similar, comparable areas possible. One way to make geographical areas comparable, is to assign geographical areas with about the same population distribution to the same group (Van Riper, Horne & Thomas, 2009). In the current study the *degree of urbanization*, which is partly dependent on the population distribution, is used. The *degree of urbanization* is an indication of the scale of human activities based on environmental density (CBS, n.d.). The higher the density, the higher the *degree of urbanization*. However, the highest *degree of urbanization* is indicated with the lowest number 1 out of 5. Another way to define comparable areas is based on *income* (Thorsnes, Alexander & Kidson, 2015). The reason why this variable is taken into account is because the income of people in a specified area may determine the availability of luxury amenities, such as cinemas or restaurants (Rigolon & Németh, 2018). Therefore, the variables of interest to make comparisons possible and filter similar geographical areas are the *degree of urbanization* and the level of *income*.

2.1.2. Data sources

CBS gathers about the proximity of different amenities in the Netherlands given the exact residential addresses of inhabitants (CBS, 2021a). An amenity is a location, such as a building, terrain or space, that can be visited by people. Different categories of amenities can be subdivided in more

specific amenities. In addition to the *average distance* to amenities, data about the number of amenities within a certain range is registered. The *number of amenities* within a certain range is the average number of amenities per person on average in a specific geographical area (CBS, n.d.). An extensive list of the categories and subcategories, list of ranges and (missing) values can be found in Appendix I. The data originate from 2020, as these data are the most recent open data available.

To compose a concise dataset, CBS gathers data using different data sources (CBS, n.d.). To determine the addresses of inhabitants, the Municipal Personal Records Database is used. Addresses of the different amenities are provided by the National Institute for Public Health and the Environment, Netherlands Institute for Health Services Research, LOCATUS, National Childcare Register, Service Execution Education, Royal Dutch Ice Skaters Association, Museums Association, Association of Theater and Concert-hall directors, Association of Dutch Music Venues and Festivals and Founding Bibliotheek.nl. Specific coordinates of the addresses are retrieved using data of the Address Coordinates of the Netherlands. Dutch translations of these institutions, as well as abbreviations can be found in Appendix IV.

Besides information about specific locations, other information about geographical objects is collected (CBS, 2021a). Road information is collected using data of the Department of Waterways and Public Works. These data are used to indicate where paved car ways are located. Furthermore, the data of RWS is useful to calculate distances from addresses to amenities. Other geographical information is provided by CBS themselves, including shapefiles containing geometries of municipalities, neighborhoods and boroughs in the Netherlands.

2.1.3. Preprocessing

First, the amenities dataset is loaded as a shapefile with the described variables of interest as different columns. From big to small, the geographical areas include the municipality, neighborhood and borough. The borough can be defined as a homogeneous area based on build environment or social economic structure (CBS, 2021a). A neighborhood consists of multiple boroughs and can be defined by unique land use purposes. The municipality is the biggest geographical area consisting of multiple neighborhoods (CBS, 2021a). The different geographical areas can be distinguished in the dataset by their unique identification code. All unique identification codes of neighborhoods and boroughs refer to the municipality and/or neighborhood they belong to (CBS, 2021a). The data about income is joined separately by these unique identification codes to obtain the correct values. For simplicity reasons and because only a small number of people live on the water, the water bodies are not taken into account. The water bodies have value ‘yes’ for the column *H2O* and could be filtered and deleted from the data. Furthermore, columns containing only ‘0-values’ were deleted from the dataset, as they do not contribute to visualizing data.

Hereafter, the coordinate reference system has been transformed to EPSG:4326 (EPSG Geodetic Parameter Dataset, 2022). Furthermore, polygons are simplified with a 0.05 proportion of points to retain to be able to visualize the data using Leaflet maps (Cheng, Karambelkar & Xie, 2022). The simplification of polygons by using the function ms_simplify() has been carried out to reduce loading time when using the dashboard (Teucher & Russell, 2021). It is assumed that very precise boundaries are not necessary for visualizing data for citizens. Another spatial operation carried out is to calculate the centroid of each polygon by using the functions st_centroid() and st_coordinates() (Pebesma, 2018).

Besides names, postal codes can be used to identify a specific area in the Netherlands. The postal codes dataset is also provided by CBS (CBS, 2021b) and are joined to the CBS-dataset about

amenities using the unique identification code for geographical areas. This way, the geographical information, as well as *distances* to and *number of amenities* can be related to *postal codes*. The postal code data originate from 2020 to make sure all observations line up correctly.

Next, an indication of income is calculated using quantiles of the percentage of households with an income under or around socially minimum income. The quantiles are used to assign a municipality and neighborhood to a group. The benefit of using quantiles is that each quantile consists of an approximately equal number of areas. The boundaries of different groups represent percentages of households living under or around social minimum wage. Geographical areas are assigned to group 1 if the percentage of households living around social minimum wage is under 3,9% (Table 1).

Table 1: Boundaries of groups for income indication

Group	Lower boundary	Upper boundary
1	≥ 1.9	< 3.9
2	≥ 3.9	< 4.6
3	≥ 4.6	< 6.0
4	≥ 6.0	≤ 13.6

2.2. Traffic incidents; using the RWS-dataset

A dataset with information about traffic incidents underlies another dashboard in the R Shiny tool (Rijkswaterstaat, 2021). Data about traffic incidents can be used for traffic safety analysis and corresponding visualizations to drive policy decisions or neighborhood initiatives. Section 2.2.1. discusses the variables that are needed for visualizations of several subtopics. Thereafter, data sources RWS uses to obtain information are discussed. Third, section 2.2.3. explains preprocessing steps to prepare the data for visualization.

2.2.1. Variables of interest

The RWS-dataset contains different columns with information about traffic incidents from 2011 up until 2020. Some variables are mandatory to register, others are optional. Mandatory variables to register for every traffic incident include a *unique registration number*, *year* in which the accident took place, *caused damage, place of impact, the number of parties* involved and the *coordinates* of the incident. Although these data are mandatory to register, 'unknown' values have been registered which are considered as missing values. Variables describing the resulting damage of the incident and the place where objects were hit are categorical variables. The *number of parties* involved is an ordinal variable. All mandatory variables are included in the preprocessed RWS-dataset for further visualization.

Optional variables that are retained in the preprocessed RWS-dataset, are the *road situation, speed limit*, *weather conditions* and *object type*. The *speed limit* can be considered an ordinal variable, all other optional variables are considered to be categorical. More information about the specific categories of different variables and missing data can be found in Appendix II. However, because of the optional nature of these variables, around 50% of the values can be missing. These variables are nevertheless retained in the RWS-dataset, because insights into these variables can be an indication of unsafe roads and locations where most of the incidents take place. Knowledge of the number of accidents for included variables and problem areas may contribute to neighborhood initiatives and policy decisions.

2.2.2. Data sources

The RWS-dataset considers registration data of road inspectors of RWS and/or police agents who are at the location of the accident (Rijkswaterstaat, n.d.). These officers enter into their own registration system the mandatory and optional variables, such as the *location* and *object type*. As a victim of the traffic incident may die as a consequence of severe injuries, it is not immediately clear if the traffic incident should be registered as 'causing injuries' or results in 'death'. Therefore, data from CBS is used to estimate the true number of traffic fatalities based on court reports (Rijkswaterstaat, n.d.). Furthermore, the Association Scientific Research Traffic Safety estimates the number of injuries caused by traffic incidents based on registrations of Dutch Hospital Data (Rijkswaterstaat, n.d.). Lastly, data from insurers is used to get insights into minor traffic incidents or one sided incidents (Rijkswaterstaat, n.d.). Yearly, these data are updated and linked with the registration system of RWS. The comparisons are done manually and have been carried out from 2018 onwards with greater accuracy (Rijkswaterstaat, n.d.).

2.2.3. Preprocessing

The original RWS-dataset is explored and prepared for use in the R Shiny dashboard tool. The original RWS-dataset includes all information about the included variables, except for X- and Y-coordinates and number of involved parties. Furthermore, categories are registered as letters or numbers. In separate files, a description is given what the numbers and letters include. The separate files are joined to the original RWS-dataset by using different identifiers. An unique registration number of the incident (*VKL-NUMMER*) is used to join the number of involved parties to the original dataset. Another unique registration number (*FK_VELD5*) is used to join the X- and Y-coordinates to the accidents. Variables that only have number or letter values are joined using their original identification to include descriptions of categories. As information about *weather conditions* is not included in a separate file, the variable is recoded manually.

To improve the quality of the data visualizations, the *object type* and *number of parties* involved are recoded. In the original dataset, 23 categories of *object types* were taken into account. The number of categories has been reduced to 10 categories for readability considerations. The variable of *involved parties* contained initially 25 categories. However, incidents involving more than 5 parties make up only 0,5% of the total number of accidents. Therefore, the number categories has been reduced to focus on the lowest numbers of involved parties. Missing data are recoded from an empty string to 'NA' or 'unknown'. The old and new values of *object types* and *number of involved parties* can be found in Appendix III.

Next, the data have been transformed to a spatial dataset with geometry features using the function st_as_sf() (Pebesma, 2018). Columns *X-COORD* and *Y-COORD* and the Dutch coordinate reference system 'EPSG:28992' are used as parameters to the function. The transformation of the coordinate reference system to 'EPGS:4326' is similar to the CBS-dataset. Lastly, a spatial intersection operation was performed so that each traffic accident can be related to the polygon data of a municipality, neighborhood and borough. The preprocessing steps resulted in a spatial dataset including variables of interest, names of *municipality*, *neighborhood* or *borough,* as well as *degree of urbanization, geometries* and *point locations*.

3. The R Shiny dashboard tool

The study has been an iterative process to find out the best way to visualize open data to empower citizens. Feedback on different versions of the dashboard was gathered in weekly meetings with scientific researchers. Comments were given on the general layout of the dashboard, as well as the way in which the data are visualized. Furthermore, feedback of citizens of Overvecht and neighboring areas is collected during the 'day of the neighborhood'. When trying different options and selection inputs, citizens commented on the look, usability and understandability of the dashboard. Most apparent feedback from citizens of Overvecht was that the overwhelming amount of data visualized at once should be reduced. Second, citizens of Overvecht did not understand the visualizations immediately and suggested to add an explanation of what is included in the dashboard. Resulting tips received from different parties were taken into account when further developing the dashboard.

The third section of this paper focusses on what programming steps are taken to result in visualizations that are understandable for citizens. First, an explanation is given about the structure of the dashboard. The next section explains what input parameters can be selected. After that, the coding behind different plots and maps will be discussed.

3.1. Amenities

In Figure 1 an overview of the amenities dashboard is shown. It consists of a side panel (1A) to switch between different topics, such as amenities and traffic incidents as included in this paper, and health care and crime data discussed by Dekkers (2022) and Kellij (2022). The main panel of the dashboard, consists of two fluid rows with each a different function. The first row, consisting of boxes 1B, 2, 3, 4, 5 and 6 (red), functions as general insight in the selected area. The second row (6 (green), 7, 8, 9) functions as a more detailed insight in the selected geographical area based on a theme and subtheme. The different colors used in the dashboard have the purpose to indicate different components. Blue boxes indicate that the user should make a choice; orange boxes are the results of the user input; the red and green box are the general top-5 and themed top-5 respectively and correspond to colors on the map of boxes 5 and 8.

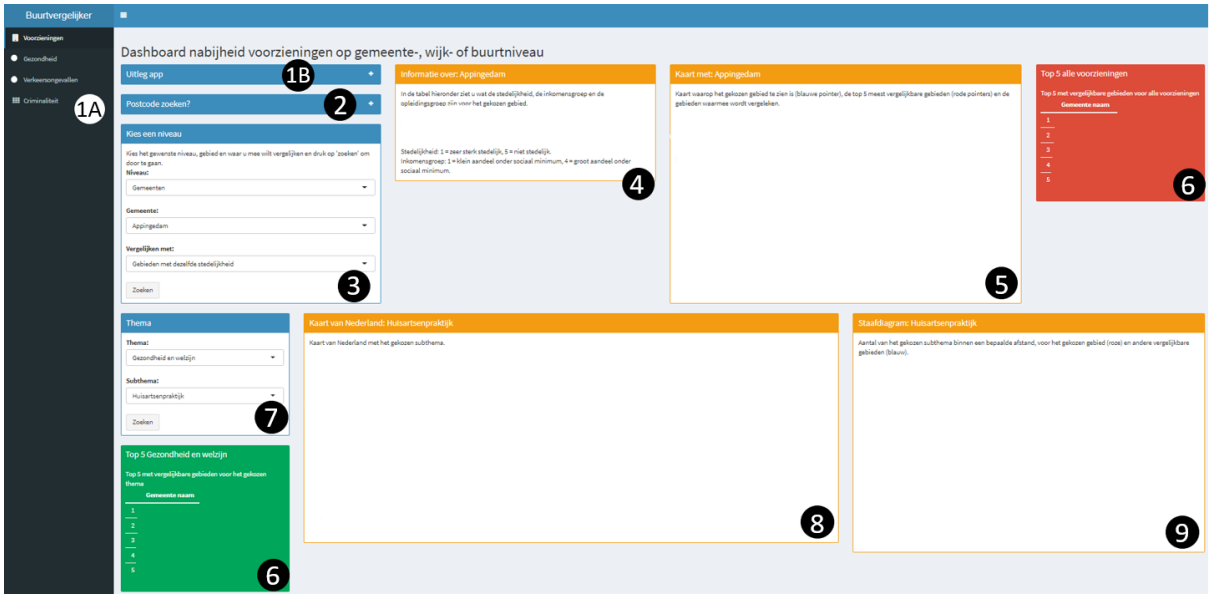

Figure 1: Overview of the amenities dashboard. Different components with different functions are numbered from 1-9.

3.1.1. Input selection

The first, but optional user input, is the postal code of the area of interest. In Figure 1, the postal code input box is marked with number 2. The box is minimized by default and can be maximized using the plus sign. The text input from the user will be normalized without spaces, capitalized and matched to a postal code in the dataset. The output is a string of text indicating the names of the municipality, neighborhood and borough that corresponds to the inserted postal code. If more neighborhoods and boroughs correspond to the inserted postal code, these are all returned in the output.

When the user is familiar with their municipality, neighborhood and/or borough, the user can specify the names in box 3. Box 3 is designed to adjust the input options based on the *level* that is selected using a conditional panel. When *municipality* is selected, the input for *neighborhood* and *borough* will be hidden. Furthermore, the input of the *municipality* is determinative for the choices that are shown in the *neighborhood* selection input. The implementation of the dependency makes sure that only unique neighborhoods that exist in the selected municipality can be chosen. The same holds for the selection of the *boroughs* based on *neighborhoods*. The filter variables of *degree of urbanization* and *group of income* can be selected to compare different geographical areas that are assigned to the same group of income or have the same degree of urbanization. However, if the *boroughs level* is selected, the *income* variable is not available due to too many missing values. The calculations will start by clicking the action button.

Lastly, in a separate box a subtheme can be selected for more specific insights. The primary themes and corresponding subthemes are predefined and the choices for the subtheme are dependent on the selection input of the primary theme. Therefore the selection input for subthemes is updated in accordance to the condition of the selected theme. Again, if the action button is clicked, the calculations will start and result in different maps and plots.

3.1.2. Filtering the CBS-dataset

A number of filters is applied to generate the correct data for visualizations. The dataset-function in the corresponding file returns for all different levels of geographical areas three objects. First, a dataset with only comparable areas is returned. The spatial CBS-dataset is temporarily transformed to a regular data frame to eliminate the *geometry* column to make indexing and slicing possible. The *degree of urbanization* or *group of income* is retrieved from the data and used to filter comparable geographical areas. Filter operations are done using base R or dplyr (Wickham, François, Henry & Müller, 2022). The dataset contains all polygons of geographical areas of the same level as the selection input from the user and is converted back to a spatial object using function st_as_sf() (Pebesma, 2018). Columns for the area *code* and name *label*, are added to the dataset. X- and Y-coordinates of the *centroid* of the selected area is retrieved using the row number of the area and inserted for all areas in a separate column. The dataset-function also returns all the observations for the selected polygon. This operation results in an one-row-data frame with all columns. Lastly, the dataset-function returns if the *income level* is known or unknown. If the *income level* is unknown, an error message will be printed stating that the comparability filter is not applied.

3.1.3. Overview of selected area

Box 4 is rendered as a table with descriptive information about the selected area. The *degree of urbanization* and the *group of income* are shown with an explanation of these variables. This information is derived from the dataset-function which returns a one-row-data frame. The columns

are renamed to short and understandable names. The information is shown in a separate box, because, it is important to know on what number comparable areas are identified.

One of the filter variables shown in box 4 is used to select comparable areas that are plotted on the map in box 5. The interactive Leaflet map uses the 'dataset' output from the dataset-function to draw the polygons on a background map. A blue marker is added to the centroid of the selected area to indicate where the area is located. Five red markers are added to the map using the centroids of the five most comparable areas as indicated in box 6 (red). The colors of the markers correspond to the color of the general top 5 box. The map can be used to gain understanding of the geographical location of the areas.

The top 5 is calculated using all variables that indicate the average distance to or the number of amenities from the 'dataset' of the dataset-function. The variables are normalized using Z-scores and the scale-function. The distance from the selected area to all comparable areas in the data frame is calculated using the function dist() and Manhattan distance. The distance matrix is sorted and merged to the initial data frame and returned as data frame with all columns of the top 5 most similar areas. The red box 6 is a table output and includes the ranking and names of comparable areas. From the resulting dataset and based on the input level, the names of comparable areas are included in the table.

3.1.4. Theme and subtheme plots

The green box 6 is another top 5 calculation. The difference between the two calculations is that the themed top 5 only takes into account the subthemes of the selected theme and not all of the themes. The themed top 5 calculates the distance of the selected area to all other comparable areas. The subthemes corresponding to the theme will be included in the subset of the dataset. Similar to the general top 5 calculation, Z-scores are calculated and a distance matrix is sorted and merged. The table to render in the R Shiny dashboard consists of a ranking and the names of the geographical areas that are most similar to the selected area. Again, names of higher *levels* are included to avoid confusion about neighborhoods or boroughs with the same name but different municipalities.

The Leaflet theme map in box 8 is an interactive map consisting of a background map and the polygons of comparable areas. The selected area is marked with a label indicating the average distance to the selected amenity and the average average distance of comparable areas to the amenity. The green markers at the centroids of the top 5 most similar areas are retrieved from the distance matrix of the above operation. The colors in the map are based on the selected subtheme and a color palette from yellow to red from R ColorBrewer-package is used (Neuwirth, 2014). The subthemes all have numeric values with an order and therefore the sequential color pallet is applied to indicate a bigger value of the variable with a darker color of the polygon (Magnuson, 2016).

The bar chart in box 9 is rendered using the 'dataset' from the dataset-function, selected theme and three predefined columns that correspond to the theme. From the dataset, the predefined columns are selected as subset and for each column the *average number of amenities* within a fixed range is calculated and added to the data frame. This information is visualized in the blue bar. The red bar is an indication of the *number of amenities* within a range of the selected area. Both datasets from the selected area and the comparable areas are bound together and form the data input for ggplot (Wickham, 2016). The bar chart is generated using the function geom_col() with X-aesthetics of the selected *subtheme*, Y-aesthetics the *number of amenities* within a fixed range and the fill-aesthetic is the *group* of comparable areas or the selected area. The insights in the differences between the selected area and comparable areas may drive change if the differences are big. The bar chart is only

shown when data about the ranges is available for the selected subtheme. Using renderUI, all boxes in the second fluid row are only shown when the second action button is clicked when selecting a theme and subtheme (Chang et al., 2021).

3.2. Traffic incidents

In Figure 2, the general overview of the traffic incidents dashboard is shown. The structure of this dashboard is similar to the dashboard of amenities. The blue box (A) indicates that an action of the user is needed for visualization. Box U is also colored blue, because it contains an explanation of the subthemes that can be selected in box A. The red boxes (B, C, D) give a general insight in the *number of traffic incidents* and the *degree of urbanization*. The second row (E, F), marked with a separate title, will show information about a *selected theme*. This row is only shown if accidents have actually happened. The third row (G, H), has green boxes and makes the comparison between areas possible. In the next subsections, each component will be explained.

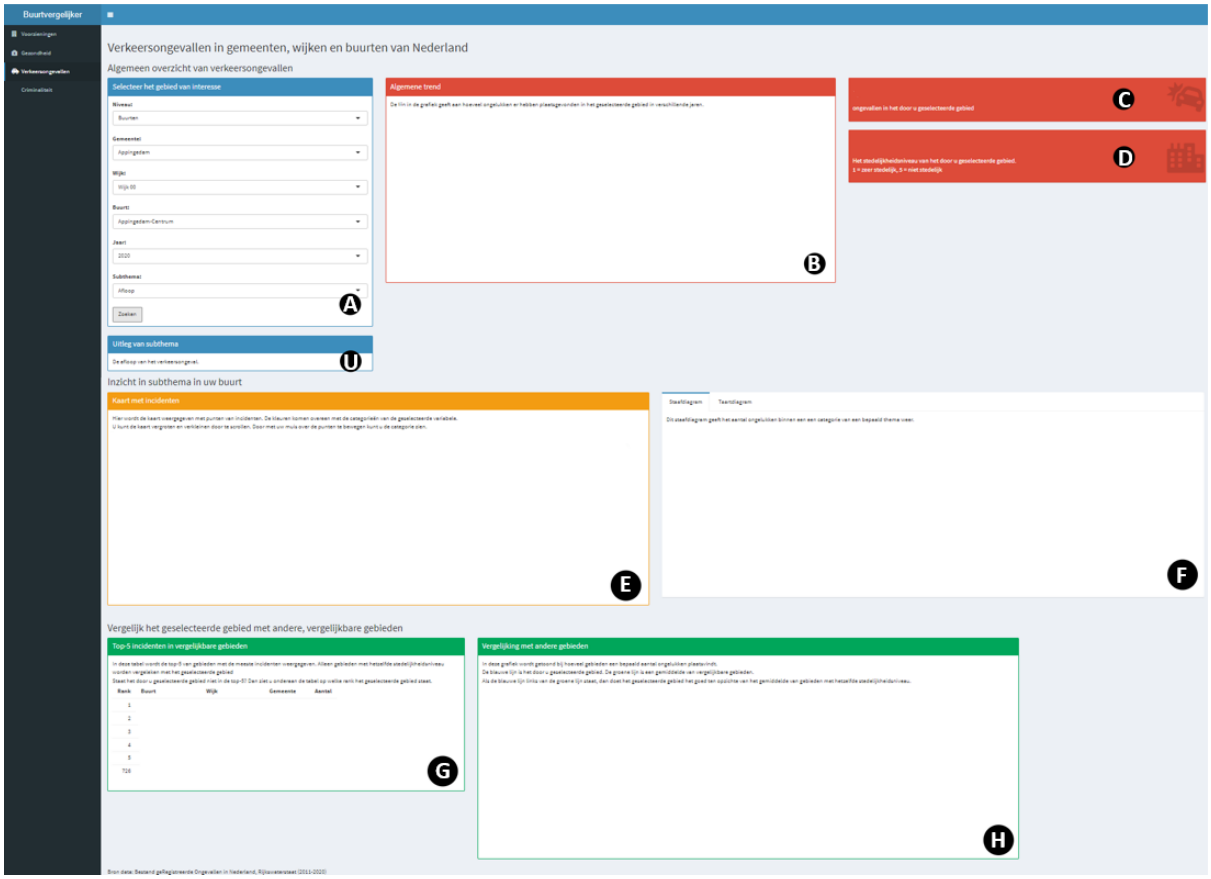

Figure 2: Overview of the traffic incidents dashboard.

3.2.1. Input selection

Blue box A forms the basis of the dashboard, because the input for *area*, *year* and *subtheme* of interest is selected. Similar to the amenities dashboard, the selected *level* is determinative for the required input parameters using conditional panels and unique choices. When selecting the *level* 'municipality', only the name of the municipality will need to be inserted. When selecting 'boroughs', the name of the municipality, as well as the neighborhood and borough need to be selected. Furthermore, it is possible to select a year of interest from 2018 up until 2020. The subthemes reflect the included categorical or ordinal variables as indicated in Appendix II.

3.2.2. Filtering the RWS-dataset

Before visualizations and other outputs can be generated, the data are filtered based on selection inputs. First, the *selected polygon* is filtered from the CBS-dataset using its input *name* and *level*. This will enable visualization of the polygon lines on the Leaflet map. Furthermore, all points in the selected area are stored in a new spatial data object. The data is filtered on the selected *level*, *year* and the *name* of the area. Third, the *number of incidents* is calculated for each unique geographical area. This is done by grouping the data by the area name and count()-function (Wickham, François, Henry & Müller, 2022). The operation results in a data frame with the *name* of the area and a column with the *counts*. Only the number of the count column is retrieved. The *degree of urbanization* is retrieved from the CBS-dataset when filtering on the name of the selected area. The *number of incidents* and *degree of urbanization* are shown in value boxes as indicated by letters C and D in Figure 2. Knowledge is this number is necessary, because this drives the comparison in the third row (box G, H) of the dashboard.

3.2.3. Trend line from 2011 to 2020

Because the number of incidents over several years is needed to plot the trend line, the returned value from the previously explained filter-function cannot be used. Therefore the RWS-dataset is filtered again on level and name of the selected area. These data, grouped by year, are counted and used as input for the ggplot-function. The aesthetics of the function geom_line() are the *years* as factors and the *number of incidents*. The trend line is colored red to match the general theme of the traffic incidents dashboard. The resulting plot shows a trend line of the *number of accidents* that happened between 2011 and 2020. Knowledge about the trend, together with experiences and perceptions of citizens, can drive the policy decisions and neighborhood initiatives to make the neighborhood safer and reduce the number of occurring accidents.

3.2.4. Specific insights in selected theme

For each subtheme, an unique color palette is defined using the ‘Set3’ or ‘Blues’ palettes from the R ColorBrewer-package (Neuwirth, 2014) and the number of categories of the subtheme. The *speed limit* and *number of involved parties* are considered to be ordinal variables and therefore the color palette is ‘Blues’ (Magnuson, 2016). All other variables are categorical values for which ‘Set3’ is used (Magnuson, 2016). The underlying data to plot on the Leaflet map is the *point-dataset* from the previously discussed filter-function. The points are drawn in different colors corresponding to distinct categories of the selected subtheme. The polylines of the selected polygon are added to the background map as a boundary. The map shows the location of the traffic incidents in the *selected area* and *year of interest*.

The map is an interactive map which makes zooming and dragging of the map possible. Zooming into a particular spot results in a more detailed look. Because the background map is added to the interactive map, the location of the accidents can be related to recognizable places in the real world. When an area is crowded with points on the map, it is an indication of a dangerous traffic situation and may need further investigation. Changing the *subtheme* changes the colors on the map, but the points remain in the same location. It can then be checked if the high number of incidents in a particular location may be caused by the *road situation*, *speed limit* or *weather conditions*. Other subthemes included give insights in the description of the accident, such as the *object type* of the involved party, *what damage is caused* to the involved parties and *where the involved parties are hit*.

The bar chart of box F counts the number of occurrences of the subtheme, because it is not easy to determine what category of the subtheme is most apparent in the map (Kirk, 2012). The number of occurrences is counted using the count-function (Wickham, François, Henry & Müller, 2022). The geom_col-function from ggplot (Wickham, 2016) is used to create a bar chart, with the *subtheme* as X-aesthetic and fill-aesthetic and the *number of occurrences* as Y-aesthetic. The exact counts are added above the bar chart as a clear indication on small screens as well. The colors have the same color palette as the interactive map of box E (Magnuson, 2016). Complementary to the bar chart in box F and based on the same information, a pie chart is plotted in another tab of box F. The pie chart is colored with the same color palette as the map and the bar chart to create consistency throughout the dashboard (Magnuson, 2016). The pie chart adds to the bar chart the information of percentages and not only the exact number of occurrences (Kirk, 2012).

3.2.5. Comparability

The third row of the traffic incidents dashboard makes a comparison between different areas possible. First, the underlying data is filtered on *level* and *year* of interest. Then, comparable areas are selected based on the same *degree of urbanization* as indicated in box D. The data of the comparable areas is grouped by the selected *level*. Then, the *number of occurrences* for each comparable area is counted. However, if no accidents happened in the selected area, the name will not occur in the RWS-dataset. Therefore, all unique area names are retrieved from the CBS-dataset. Using the anti_join-function (Wickham, François, Henry & Müller, 2022) names of areas that did not have to deal with traffic incidents are added to the RWS-dataset with '0' as value. Then, the data is sorted in descending order and the *row number* is added as ranking in a new column. If the selected area does not occur in the top 5, the *name* of the selected area, as well as *rank* and *number of incidents* is added at the end of the table. Although the *number of accidents* is the same as in box C, it makes a fast comparison between areas possible. Furthermore, the ranking gives an indication of how many areas are doing worse.

What box G cannot indicate, is the number of incidents that occur in areas that are neither in the top 5 or the selected area. Therefore a histogram in box H is included to give insights in the *distribution of the number of incidents* in comparable areas (Kirk, 2012). The histogram is plotted using ggplot (Wickham, 2016) with the dataset of 'all areas' and their corresponding *number of incidents*. The *number of incidents* is indicated on the X-axis and the *number of areas* that have dealt with this number of incidents is indicated on the Y-axis. As an indication of the *average number of accidents* comparable areas deal with, a green vertical line is added to the histogram with the number of incidents labeled. The blue vertical line is the *number of accidents* the selected area had to deal with. It then becomes clear if the selected area has to deal with a similar number of traffic accidents as comparable areas. If negative differences can be identified based on the histogram, it may drive the municipality or citizens to initiate projects to create a safer neighborhood.

4. Use of the tool for Utrecht Overvecht

In the upcoming sections, it will be discussed how the R Shiny dashboard can be used to obtain knowledge about the topics of 'amenities' and 'traffic incidents'. The case that will be discussed is about a family with children who are interested in the proximity of primary schools. The family recently moved to a new neighborhood and they have the feeling that their home is further away from primary schools then in the other neighborhood. Furthermore, the family is interested to see if the roads in the new neighborhood are safe for children and if the number of accidents is limited. All needed actions from the family and the output they will see, are explained in the following sections.

4.1. Distance to and number of primary schools

The family does not know for sure how their neighborhood and borough are called. Therefore, they search for the names of the areas by inserting their postal code in box 2. The family recently moved to the area with postal code 3561 AC. The area is located in Utrecht and the neighborhood is called Overvecht. Furthermore, the family finds out that the borough they live in is called Wolga- en Donaudreef.

Now that the family is familiar with the names of the area, they select the corresponding municipality and neighborhood in box 3 to get a general overview of the neighborhood. They decide to compare Overvecht to areas with the same *degree of urbanization* to make sure that apples are not compared to oranges. How the family submitted the neighborhood information in the boxes is shown in Figure 3 and Figure 4. When hitting the search button, the first visualizations in the boxes 4, 5 and 6 (red) are shown.

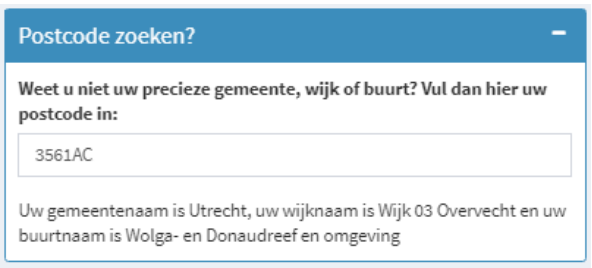

Figure 3: Box 2. Output based on postal code.

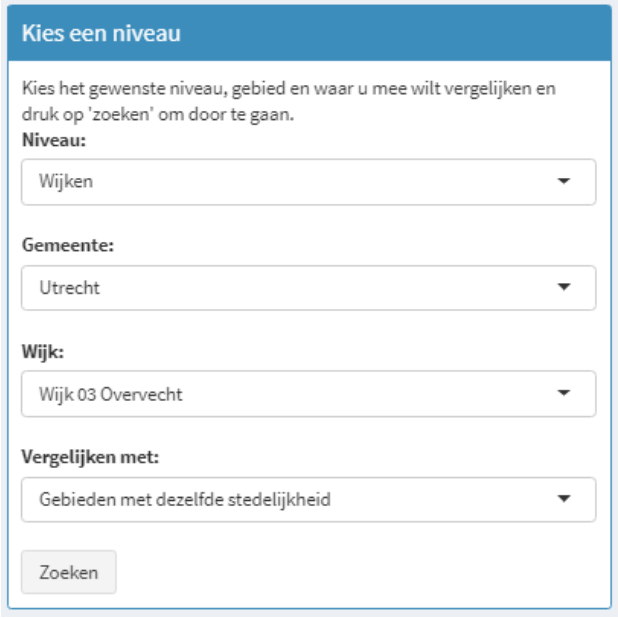

Figure 4: Box 3. Input selection of Utrecht Overvecht based on degree of urbanization

The input of box 3 results in a table, map and top 5 of similar areas based on all amenities and *degree of urbanization*. The output of box 4 tells the family that they live in a highly urbanized neighborhood. However, the group of income is quite high, which indicates that a high percentage of households deal with wages under or around social minimum. Nevertheless, only the *degree of urbanization* is taken into account for further comparisons and the *group of income* is ignored in further analysis.

Then the family takes a look at the map of box 5 and read the explanation above the map. The explanation makes clear that the blue pointer is Overvecht and that other red pointers are comparable neighborhoods based on *distances to all amenities* and *degree of urbanization*. The output of the general top 5 is shown in Figure 5. Generally, based on all amenities in the database, three neighborhoods in Utrecht are similar to Overvecht. The North-West, South and South-West of Utrecht seem to have the same distance to amenities on average as Overvecht. Two neighborhoods in Haarlem and Leiderdorp share about the same average distance to amenities as Overvecht. These neighborhoods are Europawijk and Wijk 00 respectively. The top 5 similar neighborhoods are also shown on the map in Figure 6 and makes clear to the family that Haarlem and Leiderdorp are quite far away from Overvecht.

Top 5 alle voorzieningen

Top 5 met vergelijkbare gebieden voor alle voorzieningen

	Wijk naam
1	Wijk 02 Noordwest (gemeente Utrecht)
2	Wijk 07 Zuid (gemeente Utrecht)
3	Wijk 08 Zuidwest (gemeente Utrecht)
4	Europawijk (gemeente Haarlem)
5	Wijk 00 (gemeente Leiderdorp)

Figure 5: Box 6 (red). General top 5 comparable neighborhoods to Overvecht.

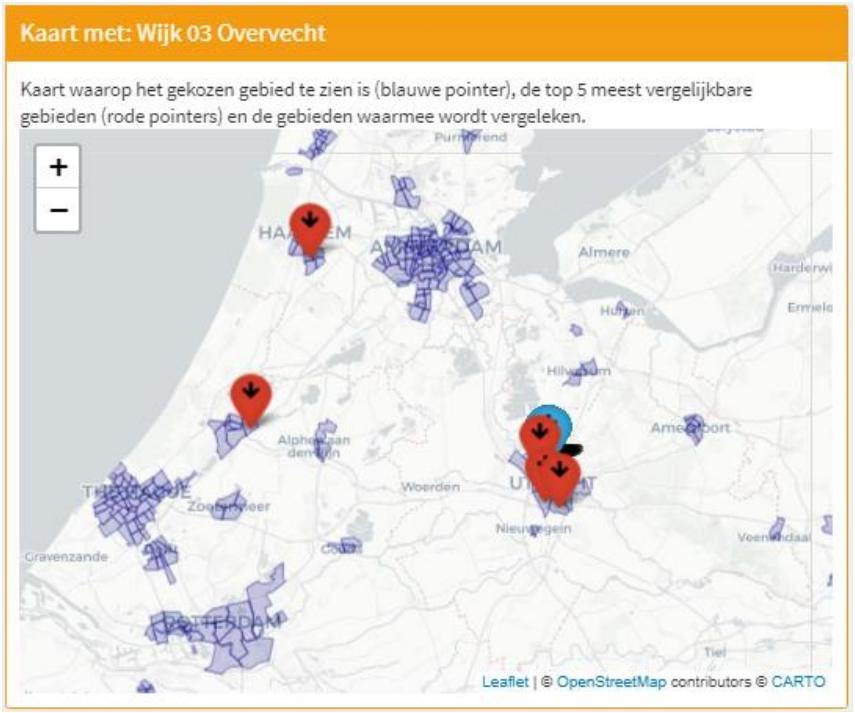

Figure 6: Box 5. Utrecht Overvecht (blue pointer), top 5 most comparable areas based on all amenities (red pointers) and comparable areas based on degree of urbanization.

A more detailed look into Overvecht is obtained when selecting 'education' as theme and 'primary schools' as subtheme. As shown on the label, in Overvecht, the average distance to a primary school is 0,5 kilometer. This is exactly the same distance as the average distance that citizens in highly urbanized neighborhoods need to travel to primary schools.

From all highly urbanized neighborhoods, only five are identified as most similar to Utrecht Overvecht. These are highlighted by the green pointers on the map in Figure 7. The names of the similar areas are summed up in a ranking in box 6 (green). As Figure 8 indicates, only one neighborhood in Utrecht is similar to Overvecht. This means that when having a look at primary schools specifically, the South-West and North-West of Utrecht and Europawijk of Haarlem are no longer comparable to Overvecht. Other neighborhoods outside of Utrecht that do compare to Overvecht based on the average distance to primary schools and degree of urbanization, are Wijk 00 in Leiderdorp, Heemskerk-Dorp in Heemskerk, Centrum in Vlaardingen and Hofland, Oosterwijk and Zuidbroek in Heemskerk.

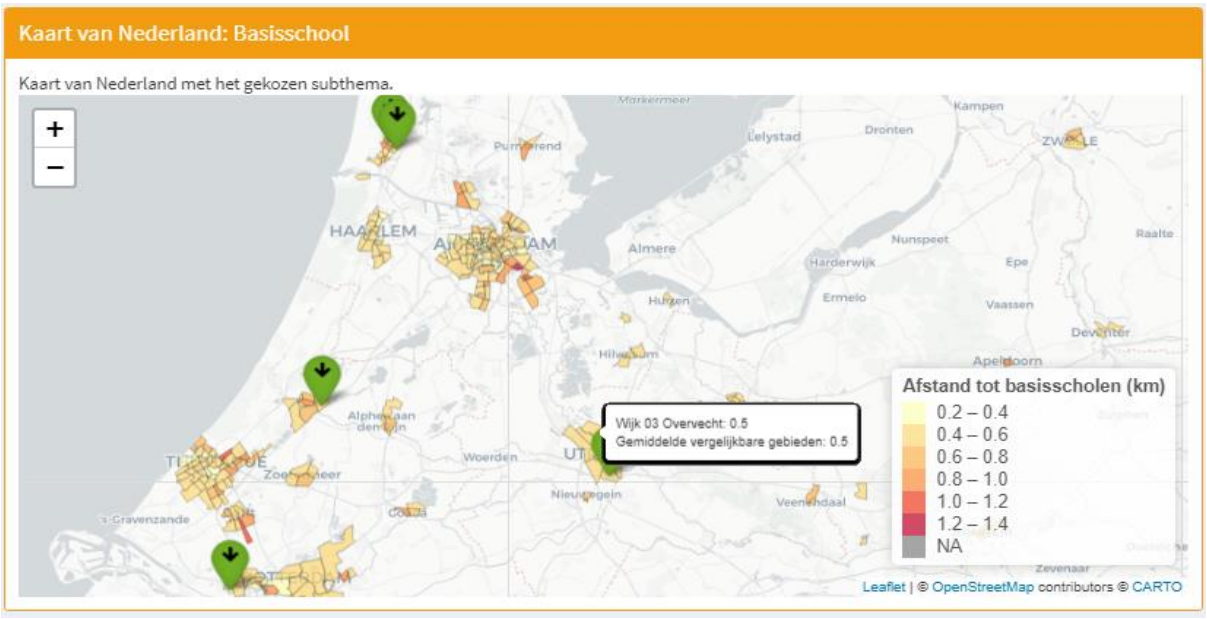

Figure 7: Box 8. Average distance to primary schools in Overvecht and the top 5 most similar neighborhoods.

Figure 8: Box 6 (green). Top 5 most similar neighborhoods to Overvecht based on average distance to primary school.

Lastly, the input of the family for Utrecht Overvecht and primary schools results in a bar chart indicating the *number of amenities* within a range of 1, 3 and 5 kilometer. The bar chart shows that in Overvecht, less primary schools are available for each range compared to the average number of primary schools in comparable, highly urbanized neighborhoods. However, the differences between Utrecht Overvecht and other highly urbanized neighborhoods seem to be small. Particularly within a range of 1 and 3 kilometer, the other highly urbanized neighborhoods have only one or two primary schools more on average. The bar chart is visualized in Figure 9.

Figure 9: Box 9. Bar chart number of amenities in Overvecht and average in comparable neighborhoods.

4.2. Traffic incidents in Overvecht for different road situations

Now that the family knows that the average distance to primary schools is the same as the distance other citizens in highly urbanized neighborhoods have to travel, the family is interested in the road safety in Utrecht Overvecht. The municipality and neighborhood names they retrieved from the postal code lookup in the amenities dashboard are selected again in the traffic incidents dashboard. The family is specifically interested in the number of accidents in 2020 and the road situation.

First, a line plot is shown in box B to gain knowledge about the trend of the *number of accidents* in Utrecht Overvecht between 2011 and 2020. The trend line shown in Figure 10 indicates a steep drop in the number of incidents in 2014. If this is considered an outlier, the general trend is that the number of incidents declines over the years. The *total number of accidents* in 2020 in Overvecht is 264, which is shown in box C. Box D indicates the *degree of urbanization* which is underlying the selection of comparable areas. Both box C and D are displayed in Figure 11.

Figure 10: Box B. Trend line of number of incidents Overvecht.

Figure 11: Box C and D. Indication of number of incidents and degree of urbanization of Overvecht.

Box E shows a map with colored points which indicates where the incident have happened and what category of the *road situation* corresponds to the particular incident. As can be seen on the map in Figure 12, in the red circles [*added*] no accidents happened in the West and North of Overvecht. The least amount of points is colored in the turquoise, however, that is not easy to tell using only the map. The information in the map is therefore complemented with the bar chart and pie chart. The bar chart in Figure 13 is in line with the reasoning above that the turquoise color is the least apparent. Only 10 incidents have happened in a turn of the road. The most incidents happen on the straight road. The pie chart in Figure 14 complements the bar chart in a way that the percentages of the category relative to the total number of incidents are shown. Resulting from the pie chart, only 3,8% of the incidents happens in a turn against 29,9% on the straight.

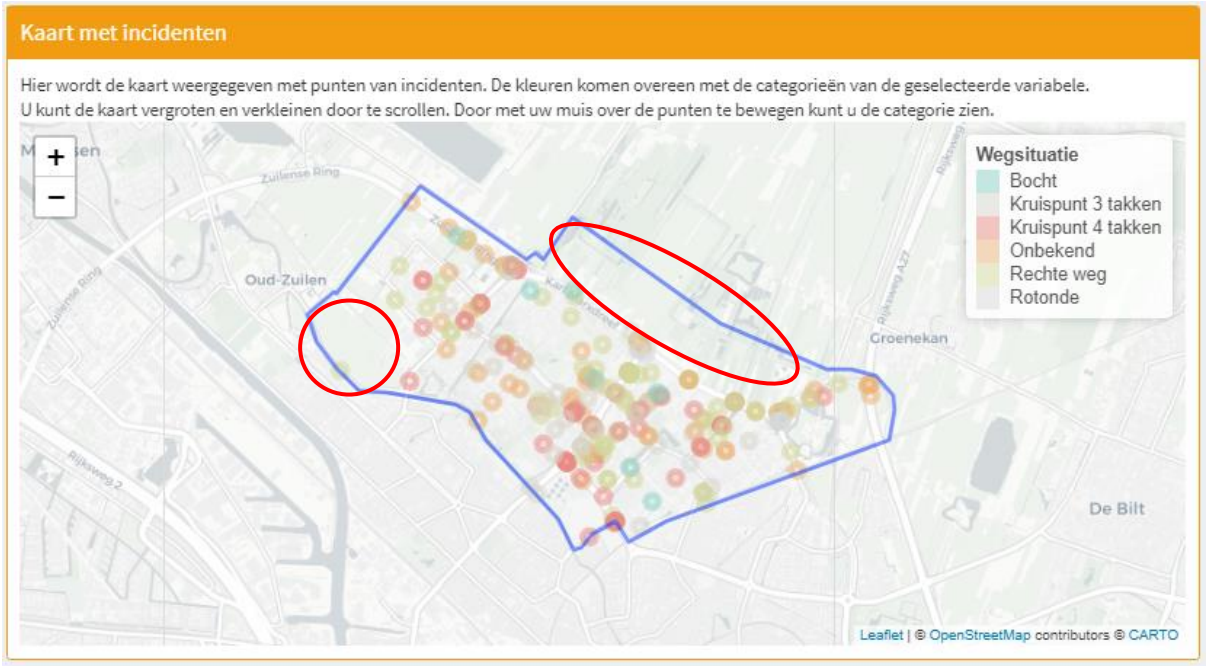

Figure 12: Box E. Map of incidents in Overvecht colored by road situation in 2020.

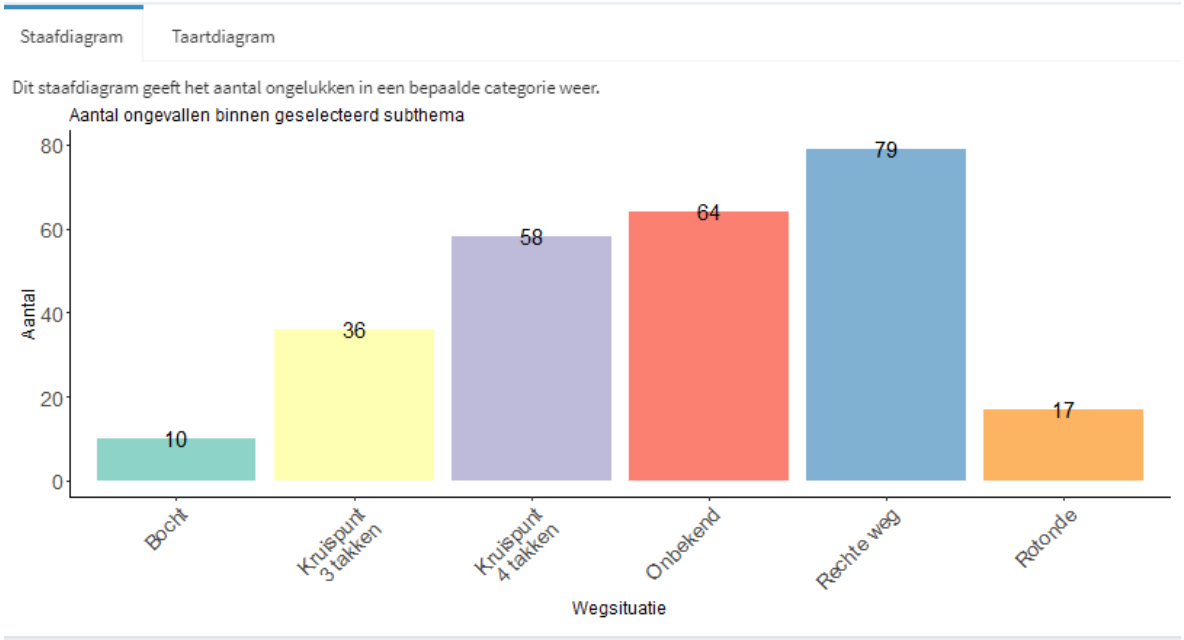

Figure 13: Box F (bar chart). The number of incidents per category of road situation in 2020 in Overvecht..

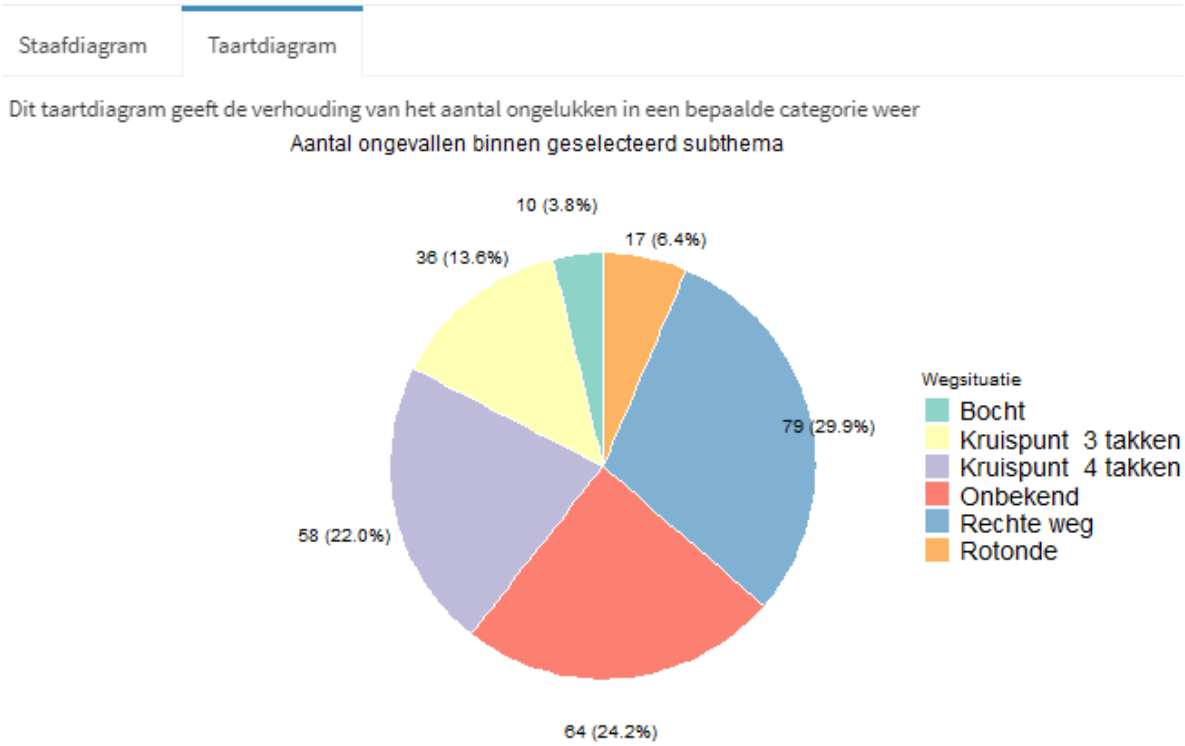

Figure 14: Box F (pie chart). The number of incidents per category of road situation in 2020 in Overvecht.

Second to last, a table with the top 5 neighborhoods that deal with the most traffic incidents is shown. The ranking in Figure 15 shows that among highly urbanized neighborhoods, Overvecht is ranked 14th. Rotterdam however, leads the ranking with five neighborhoods that deal with a high number of incidents ranging from 715 to 535. Now, the family concludes that a total number of incidents of 264 in comparison to Rotterdam is not so bad.

Top-5 incidenten in vergelijkbare gebieden

In deze tabel wordt de top-5 van gebieden met de meeste incidenten weergegeven.
Staat uw wijk niet in de top-5? Dan ziet u onderaan de tabel op welke rank het geselecteerde gebied staat.

Rank	Wijk	Gemeente	Aantal
1	Prins Alexander	Rotterdam	715.00
2	IJsselmonde	Rotterdam	624.00
3	Charlois	Rotterdam	597.00
4	Feijenoord	Rotterdam	544.00
5	Noord	Rotterdam	535.00
14	Wijk 03 Overvecht	Utrecht	264.00

Figure 15: Box G. Top 5 neighborhoods with the highest degree of urbanization and number of incidents in 2020.

The ranking is invigorated by the histogram which shows the distribution of the number of incidents among highly urbanized neighborhoods as selected by the family. The green line and label in the histogram in Figure 16 are an indication of the *average number of traffic incidents* in highly

urbanized neighborhoods. In 2020, on average 68 accidents have happened in similar neighborhoods to Overvecht. However, as the blue line indicates, Overvecht has to deal with more traffic incidents compared to the average of similar neighborhoods. This result is contradicting the positive feeling of the family from the top 5 output.

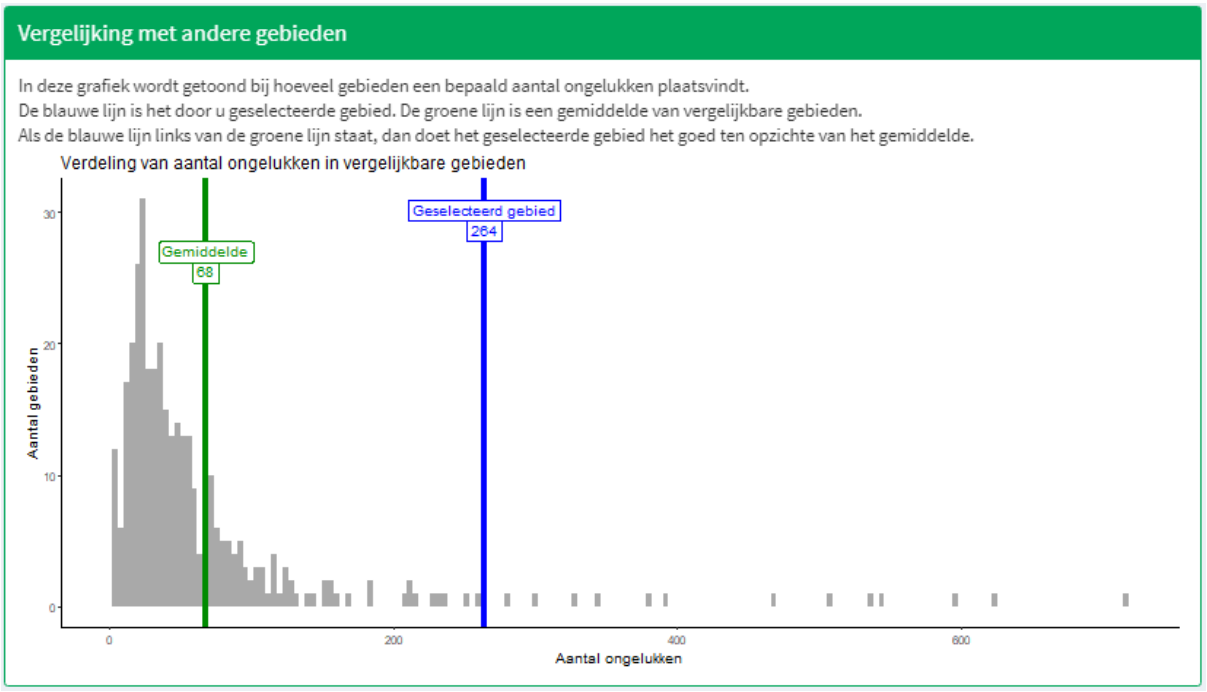

Figure 16: Box H. Histogram of the distribution of incidents for highly urbanized neighborhoods in 2020.

5. Discussion

5.1. Quality of the original data

The data that are put into the R Shiny dashboard tool determine part of the quality of the visualizations. The original CBS-dataset used (CBS, 2021) limits the number of variables taken into account in the dashboard, because certain variables deal with a lot of missing values. This limitation results in not gaining insights in these variables, even though they could be of interest for citizens. One example is the distance to a fire fighter station. In case of an emergency situation, a close by fire fighter station may be able to prevent further damage if it can be present at the location in a short amount of time. If the fire fighter station is far away, citizens may want to establish their voluntary fire fighter organization. Another example of a variable that could have been of interest, is the filter variable of *income* on *borough level*. Because of too many missing values the *group of income* could only be assigned to a small number of boroughs. If these values were known, it could contribute to knowledge about the suitability of establishing an amenity in the borough and the support base for the amenity.

A second limitation of the study is that the average distance to amenities is calculated using paved car roads (CBS, 2021a). This results in the fact that bicycle lanes and pavements are not taken into consideration. As the Dutch citizens often travel by bike and foot, it would have made sense to include this calculation and determine the shortest distance. Furthermore, some citizens may not have the resources to travel by car even if this is the shortest road to take.

Besides the discussed limitations of the original CBS-dataset, thoughts have put into maintaining quality of the dashboard. This has been done by using open data for both postal codes and neighborhood statistics of the same year, namely 2020. These data are the most recent data openly available to citizens. By combining postal code information and neighborhood statistics of the year 2020, the correct geographical area is related to the inserted postal code in the dashboard. Furthermore, data from the same year take into account the same geographical reclassification from previous years. Thus, the names of the geographical area and postal code correctly line up and confusion about the name of the area compared to previous years is avoided.

The limitation of reclassifications of geographical boundaries also involves the original RWS-dataset (2021). Therefore, the resulting trend line of the number of incidents may differ from the actual number of incidents happened in the area. When the reclassification of the geographical area of interest results in a bigger surface of the area, this may result in an overestimation of the number of incidents that happened in a previous year in the area of interest. Likewise, the reclassification may lead to an underestimation of the number of incidents in previous years in the geographical area that became smaller. To contain the error related to geographical reclassifications, only a limited number of years can be selected as input. Possible changes in boundaries are then limited as much as possible, yet still handing users of the dashboard to select other years to detect changes in time within a certain category.

Second, the quality of the RWS-dataset before 2017 is limited. Officers were not very precise in registering the exact location where the incident happened. Sometimes only the name of the road or road polygon was registered. Analysts manually checking the data had to fill in the X- and Y-coordinates of the center of the object (Rijkswaterstaat, 2021). As can be imagined, this is not a precise location and will not represent the data well on the map. Multiple registered accidents with loosely interpreted location coordinates can therefore result in a lot of incidents seemingly happening at the same location. However, the inaccuracy of the registered location does not impact the count of the incidents and therefore the bar chart, pie chart, histogram and table do not suffer from this issue.

5.2. Usability of the dashboard

The dashboard tool is initially designed for citizens who are highly active in the neighborhood and interested in data. A prerequisite for successful use by citizens, is that citizens have a particular level of knowledge about visualizations and interpreting data (Montes & Slater, 2019; Yoon & Copeland, 2019). However, at the 'day of the neighborhood' in Utrecht Overvecht became clear that citizens had difficulties in processing the amount of data correctly. This feedback resulted in loading different parts of the dashboard at different input actions to reduce the amount of information shown at once. A mediator can help citizens to correctly interpret the visualizations of themes the citizens are interested in. The explanation can help to avoid different or faulty interpretations of the data and makes sure every citizen is on the same page regarding the same subject.

5.3. Ethical and legal considerations

When not a lot of citizens live in a particular area, results will not be published to protect the citizens' privacy (CBS, n.d.). When less than 10 people live in an industrial area or at the country side, it could be possible to identify the precise measurements of the distance to amenities. In this case, the results of *average distance* and the *number of amenities* within a certain range are not published. However, collecting data about residential addresses is not unethical, as the Municipal Personal Records Database contains all addresses of all Dutch inhabitants. These addresses are already known and are mandatory for municipalities to report to the governmental institutions (CBS, n.d.). Besides, the exact addresses are not reported in the dataset, so there is no infringement of privacy.

As well as the CBS-dataset as discussed above, the RWS-dataset also has to deal with privacy regulations which impacted the content of the dataset. Due to privacy law (AVG), data about victims of traffic incidents is not publicly available. This is done to protect the victims from being identified based on the data. The privacy regulation in the Netherlands also requires several variables to be removed from the open dataset (Rijkswaterstaat, 2021). This is true for variables containing vehicle information such as number plate, measurements, insurance and periodic vehicle inspection information. The excluded variables are not particularly interesting for visualizations and therefore the exclusion of these variables did not impact the completeness or quality of the dashboard.

5.4. Future research

In the future it may be interesting to visualize more open data to empower citizens to start their own community initiatives. Among other topics, citizens may be interested in data about noise pollution caused by trains or airplanes to strengthen their standing point. Furthermore, open data is available to investigate the suitability of areas for sustainability measures such as solar panels or wind turbines. These kind of data require other visualizations in a dashboard for citizens. Therefore, it is important that citizens are involved in the project in various ways. Citizens should be able to voice their opinion on subjects of their interests. Furthermore, the chosen suitable visualizations should be in line with their data literacy. Citizens can elaborate and give feedback on the understandability of the dashboard and visualizations if they are involved in the iterative process of designing the dashboard.

A second option to dive into in future research is how to avoid the limitations regarding geographical reclassifications. Several analysis methods such as area interpolation can be used to see if an incident in 2019 happened in a particular area is still called the same in 2020. It could also be helpful if people could insert their postal code and then a range from the centroid of this postal code

area is calculated and marked. From the centroid of the postal code area, a range can be used to identify the number of traffic incidents.

Third, the loading speed of the dashboard and visualizations can be improved. Options to load leaflet maps faster, is to draw a background map only once and then projecting polygons and points on the pre-drawn map instead of drawing the map again each time a new input parameter is selected. The dashboard may also be improved if different visualizations are displayed whenever they are calculated and not all at once when every individual calculation is performed. This could make sure that citizens or the mediator can already look at the visualizations that take less time. When the visualizations are then interpreted and discussed, the more calculation intensive visualizations are shown and ready to be interpreted.

6. Conclusion

The current study aimed to empower citizens with openly available data through visualizations in a R Shiny dashboard tool. The tool contributes to the democratic process as citizens have the right to be properly informed about policy decisions of the government (Ruiter et al., 2017). Furthermore, the data available and understandable for citizens can help them to initialize projects in the neighborhood to improve the living environment. Data about amenities and traffic incidents were not yet visualized, easily accessible and understandable for citizens in already existing online dashboards (Allecijfers.nl, n.d.; Incijfers.nl, n.d.). Therefore, the current study tried to bridge this gap by fulfilling the need to have insights in specific topics and comparisons between different geographical areas. The code to create the dashboard is published on GitHub[2]. Open code can contribute to the further processing of open data in visualization dashboards. As researchers may want to visualize other open data, the open code can be used as reference.

The main focus of the study was how open data can be visualized as accurately and comprehensible as possible for citizens in the Netherlands. The data as presented in the dashboard tool help to obtain insights in the neighborhood citizens live in, as well as to compare the neighborhood with other areas in the Netherlands. Different visualizations are used to shine a light on different aspects of the data in a proper way. In the amenities dashboard, maps are included as an indication of the location of comparable areas and to visualize distances to amenities with corresponding colors. The amenities dashboard also includes tables to rank areas based on their similarity or as description of the selected area. The bar chart is used to visualize the number of amenities within a certain range (Kirk, 2012). The process to the end product was an iterative process and citizens of Utrecht Overvecht as well as scientific researchers provided feedback to improve the usability and understandability of the dashboard.

In the traffic incidents dashboard, a bar chart, table and map are also used to visualize the data (Kirk, 2012). However, the purpose of the map in this case is to show the location of traffic incidents and the category it belongs to. The number of accidents within a category are counted in the bar chart and pie chart (Kirk, 2012). The purpose of the table is the same as in the amenities dashboard, namely to rank the comparable areas. The traffic incidents dashboard also includes a trend line to identify changes over time and a histogram to get insights in the distribution of the number of accidents in comparable areas.

[2] https://github.com/ImkeDekkers/Buurtvergelijker. An explanation of the structure of this repository can be found in Appendix V.

References

Allecijfers.nl (n.d.). Retrieved June 30, 2022, from https://allecijfers.nl/

Arnhemse Koerier (2022, January 10). Stem op burgerinitiatief Jeu de Boules baan Immerloo park. *Arnhemse Koerier. https://www.arnhemsekoerier.nl*

Bolt, N. (2022, April 28). Buren leggen samen natuurvriendelijke kruising aan. *Steenwijker Courant.* https://steenwijkercourant.nl

Centraal Bureau voor de Statistiek [CBS] (2021a). *Nabijheid voorzieningen; afstand locatie, wijk- en buurtcijfers 2020* [Dataset and table information]. Retrieved on June 9 2022, from https://opendata.cbs.nl/statline

Centraal Bureau voor de Statistiek [CBS] (2021b). *Kerncijfers per postcode* [Dataset]. Retrieved on 29 April, from https://www.cbs.nl

Centraal Bureau voor de Statistiek [CBS] (n.d.). *Nabijheidsstatistiek*. Retrieved on June 9 2022, from https://www.cbs.nl

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges. B. (2021). *R Shiny: Web Application Framework for R*. R package version 1.7.1., https://CRAN.R-project.org/package=shiny

Cheng, J., Karambelkar, B., & Xie, Y. (2022). *Leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*. R package version 2.1.0. https://CRAN.R-project.org/package=leaflet

Dekkers, I. (2022). *Area comparisons on municipality, neighbourhood, and borough level: Shiny App in R for open data about amenities and health* [Unpublished master's thesis]. Utrecht University.

Duurzaam Amsterdam (2022, February 3). Oost Begroot: 1,2 miljoen euro voor buurtinitiatieven. *Duurzaam Amsterdam.* https://duurzaamamsterdam.net

Echt Overvecht (2022, March 14). *Kasaidreef in Overvecht van 50 naar 30 km/h.* Retrieved on May 31 2022, van https://www.echtovervecht.nl

Echt Overvecht (n.d.). *Samen voor Overvecht.* Retrieved on May 30 2022, from https://www.echtovervecht.nl

EPSG Geodetic Parameter Dataset. Available online: https://epsg.org (accessed on June 29 2022).

Gemeente Utrecht (2017, October 20). *Welke activiteiten heeft De Versnelling in de wijk?* Echt Overvecht. Retrieved on June 1 2022, from https://www.echtovervecht.nl

Gemeente Utrecht (2019). *Samen voor Overvecht.* Utrecht: Gemeente Utrecht.

Incijfers.nl (n.d.). Retrieved June 30, 2022, from https://incijfers.nl/

Kellij, S.A. (2022). *Compare your neighborhood's amenities and crimes. Visualizing and comparing open data with R shiny on the municipal, neighborhood and borough level for citizens* [Unpublished master's thesis]. Utrecht University.

Kim, S. & Lee, J. (2019). Citizen Participation, Process, and transparency in Local Government: An Exploratory Study. *Policy Studies Journal, 47*(4), 1020-1041. https://doi.org/10.1111/psj.12236

Kirk, A. (2012). *Data Visualization: a successful design process.* Birmingham: Packt Publishing.

Luthfi, A. & Janssen, M. (2019). Open Data for Evidence-based Decision-making: Data-driven Government Resulting in Uncertainty and Polarization. *International Journal on Advanced Science, Engineering and Information Technology, 9*(3), 1071-1078. https://doi.org/10.18517/ijaseit.9.3.8846

Magnuson, L. (2016). *Data Visualization. A Guide to Visual Storytelling for Libraries.* London: Rowman & Littlefield.

Meijer, A.J., Curtin, D., & Hillebrandt, M. (2012). Open government: connecting vision and voice. *International Review of Administrative Sciences, 78*(1), 10-29. https://doi.org/10.1177/0020852311429533

Michels, A. & Graaf, L. de (2010). Examining Citizen Participation. Local Participatory Policy Making and Democracy. *Local Government Studies, 36*(4), 477-491. https://doi.org/10.1080/03003930.2010.494101

Montes, M.C. & Slater, D. (2019). In T. Davies, S. Walker, M. Rubinstein & F. Perini (Eds.), *The State of Open Data: Histories and Horizons* (pp. 274-286). Cape Town: African Minds.

Movisie (2012). *Ondersteuning burgerinitiatieven door steunpunten vrijwilligers werk*. Utrecht: Movisie.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes.* R package function 1.1-2. https://CRAN.R-project.org/package=RColorBrewer

Oosterhout Nieuws.nl (2021, Juli 7). Gemeente stelt 3 ton ter beschikking voor buurtinitiatieven. *Oosterhout Nieuws.nl*. https://oosterhout.nieuws.nl

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

Rigolon, A. & Németh, J. (2018). What Shapes Uneven Access to Urban Amenities? Thick Injustice and the Legacy of Racial Discrimination in Denver's Parks. *Journal of Planning Education and Research, 41*(3), 312-325. https://doi.org/10.1177/0739456X18789251

Rijkswaterstaat (2021). *Verkeersongevallen – Bestand geRegistreerde Ongevallen Nederland* [Dataset and code book]. Retrieved from https://nationaalgeoregister.nl

Rijkswaterstaat (n.d.). *Bronnen voor ongevallencijfers*. Retrieved on June 9 2022, from https://www.rijkswaterstaat.nl

Ruiter, E., Grimmelikhuijsen, S., & Meijer, A. (2017). Open data for democracy: Developing a theoretical framework for open data use. *Government Information Quaterly, 34*(1), 45-52. http://dx.doi.org/10.1016/j.giq.2017.01.001

Samen voor Overvecht (2022). *Resultaten wijkaanpak Samen voor Overvecht. Voorjaar 2022.* Utrecht: Samen voor Overvecht.

Smit, R. (2022, February 20). Buurtinitiatief de Buurttuin in Overhees groot success. *Soester Courant.nl.* https://www.soestercourant.nl

Teucher, A. & Russell, K. (2021). *rmapshaper: Client for 'mapshaper' for 'Geospatial' Operations.* R package version 0.4.5. https://CRAN.R-project.org/package=rmapshaper

Thorsnes, P., Alexander, R., & Kidson, D. (2015). Low-income housing in high-amenity areas: Long-run effects on residential development. *Urban studies, 52*(2), 261-278. https://doi.org/10.1177/0042098014528999

Ubaldi, B. (2019). Governments. In T. Davies, S. Walker, M. Rubinstein & F. Perini (Eds.), *The State of Open Data: Histories and Horizons* (pp. 381-394). Cape Town: African Minds.

Van Riper, D., Horne, K., & Thomas, W.L. (2009, 8-11 June). *Comparable geography: options for developing small area harmonized geographies for comparison, implications for aggregate data production* [Conference paper]. International Conference on Intelligence and Security Informatics: Richardson, Texas, United States of America.

Verba, S., Schlozman, K.L., & Brady, H. (1995). *Voice and Equality: Civic Voluntarism in American Politics.* London: Harvard University Press.

Wegdam, E. (2022, March 22). *Verslag inspiratiesessie 3.* Utrecht: Wijkplatform Overvecht.

Wessels, B., Finn, R., Sveinsdottir, T., & Wadhwa, K. (2017). *Open Data and the Knowledge Society.* Amsterdam: University Press.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis.* R package version 3.3.5. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *Dplyr. A Grammar of Data Manipulation.* R package version 1.0.8. https://CRAN.R-project.org/package=dplyr

Wilson, C. (2019). Civil Society. In T. Davies, S. Walker, M. Rubinstein & F. Perini (Eds.), *The State of Open Data: Histories and Horizons* (pp. 355-366). Cape Town: African Minds.

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

Appendix I
Description amenities variables

Table 2: Included variables of amenities dataset

Variable	Values	% missing values municipality	% missing values neighborhood	% missing values borough
General				
Name of municipality	Text	0		
Code of municipality	Text	0		
Name of neighborhood	Text		0	
Code of neighborhood	Text		0	
Name of borough	Text			0
Code of borough	Text			0
Level	Categorical: Municipalities Neighborhoods Boroughs	0	0	0
Degree of urbanization	Integer (1-5)	0	0.09	0.5
Group of income	% households under or around social minimum wage	0	8.59	31.21
Health and wellbeing				
Distance to general practice		0	0.85	4.85
Number of general practices	1, 3, 5 km	0	0.85	4.85
Distance to hospital with clinic		0	0.85	4.85
Number of hospitals with clinic	5, 10, 20 km	0	0.85	4.85
Distance to hospital without clinic		0	0.85	4.85
Number of hospitals without clinic	5, 10, 20 km	0	0.85	4.85
Distance to pharmacy		0	0.85	4.85

Variable	Values	% missing values municipality	% missing values neighborhood	% missing values borough
Retail				
Distance to supermarket		0	0.85	4.85
Number of supermarkets	1, 3, 5 km	0	0.85	4.85
Distance to daily general provisions		0	0.85	4.85
Number of daily general provisions	1, 3, 5 km	0	0.85	4.85
Distance to warehouse		0	0.85	4.85
Number of warehouses	5, 10, 20 km	0	0.85	4.85
Catering industry				
Distance to pub		0	0.85	4.85
Number of pubs	1, 3, 5 km	0	0.85	4.85
Distance to cafeteria		0	0.85	4.85
Number of cafeterias	1, 3, 5 km	0	0.85	4.85
Distance to restaurant		0	0.85	4.85
Number of restaurants	1, 3, 5 km	0	0.85	4.85
Distance to hotel		0	0.85	4.85
Number of hotels	5, 10, 20 km	0	0.85	4.85
Day-care				
Distance to day-care		0	0.85	4.85
Number of day-cares	1, 3, 5 km	0	0.85	4.85
Distance to school-care		0	0.85	4.85
Number of school-cares	1, 3, 5 km	0	0.85	4.85

Variable	Values	% missing values municipality	% missing values neighborhood	% missing values borough
Education				
Distance to primary school		0	0.85	4.85
Number of primary schools	1, 3, 5 km	0	0.85	4.85
Distance to secondary school		0	0.85	4.85
Number of secondary schools	3, 5, 10 km	0	0.85	4.85
Distance to VMBO-school		0	0.85	4.85
Number of VMBO-schools	3, 5, 10 km	0	0.85	4.85
Distance to Havo/Vwo-school		0	0.85	4.85
Number of Havo-Vwo-schools	3, 5, 10 km	0	0.85	4.85
Traffic and transport				
Distance to drive way main road		0	0.85	4.85
Distance to train station		0	0.85	4.85
Distance to important transfer station		0	0.85	4.85

Variable	Values	% missing values municipality	% missing values neighborhood	% missing values borough
Leasure and culture				
Distance to cinema		0	0.85	4.85
Number of cinemas	5, 10, 20 km	0	0.85	4.85
Distance to theme park		0	0.85	4.85
Number of theme parks	10, 20, 50 km	0	0.85	4.85
Distance to performing arts		0	0.85	4.85
Number of performing arts podiums	5, 10, 20 km	0	0.85	4.85
Distance to museum		0	0.85	4.85
Number of museums	5, 10, 20 km	0	0.85	4.85
Distance to swimmingpool		0	0.85	4.85
Distance to figure skating		0	0.85	4.85
Distance to library		0	0.85	4.85
Distance to pop podium		0	0.85	4.85
Distance to sauna		0	0.85	4.85
Distance to tanning bed		0	0.85	4.85

Appendix II
Description traffic incidents variables

Table 3: Included variables of traffic incidents dataset

Variable	**Scale**	**Mandatory**	**% Missing values**	**Values**
VKL-identification	Text	Yes	0	
Year	Integer	Yes	0	
Outcome of accident	Categorical	Yes	0	Only material damage Injury Death
Number of involved parties	Ordinal	Yes	0	0 1 2 3 4 5+
Character of the incident	Categorical	Yes	59.31	Head-on One-sided Fixed object Flank Head-tail collision Pedestrian Parked vehicle Loose object Animal
Road situation	Categorical	No	48.46	Straight way Turn Intersection 3 ways Intersection 4 ways Roundabout Straight way, separate lanes Insertion lane Slip road

Variable	Scale	Mandatory	% Missing values	Content
Maximum speed	Ordinal	No	41.28	15 30 50 60 70 80 90 100 130
Weather conditions	Categorical	No	50.17	Dry Rain Snow or hail Strong wind Fog
Object type	Categorical	No	42.23	Motorized bicycle or scooter Bicycle Motor Agricultural vehicle Object Passenger car or delivery van Truck or bus Pedestrian Train or tram
Point location: X-coordinate	EPSG:4289	Yes	0	
Point location: Y-coordinate	EPSG:4289	Yes	0	

Appendix III
Renaming scheme traffic incidents

Table 4: Indication of frequency distribution and renaming scheme for all years and levels

Variable	Old value	Percentage	New value
Number of involved parties	0	40.6%	0
	1	12.4%	1
	2	38.4%	2
	3	6.7%	3
	4	1.4%	4
	5	0.3%	5+
	6	0.1%	5+
	7	0.01%	5+
	8	<0.01%	5+
	9		5+
	10		5+
	11		5+
	12		5+
	13		5+
	14		5+
	15		5+
	16		5+
	17		5+
	18		5+
	19		5+
	20		5+
	24		5+
	25		5+
	31		5+
	36		5+
	45		5+
Object type	e-bike	7.8%	Motorized bicycle or scooter
	Bromfiets		
	Brommobiel		
	Scootmobiel		
	Snorfiets		

Variable	Old value	Percentage	New value
Object type	Bicycle	2.4%	Bicycle
	Tree Animal Light pole Road furniture Fixed object Lose object	0.4%	Object
	Pedestrian	0.1%	Pedestrian
	Unknown	42.2%	Unknown
	Delivery van Passenger car	45.5%	Passenger car or delivery van
	Tractor Agricultural vehicle Tractor and trailer	0.9%	Agricultural vehicle
	Train or tram	0.03%	Train or tram
	Motor	1.3%	Motor
	Truck Bus	1.4%	Truck or bus

Appendix IV
Translations and abbreviations

Table 5: List of translations and abbreviations

English name	Dutch name	Abbreviation
Department of Waterways and Public Works	Rijkswaterstaat	RWS
Statistics Netherlands	Centraal Bureau voor de Statistiek	CBS
Municipal Personal Records Database	Gemeentelijke Basisadministratie	GBA
National Institute for Public Health and the Environment	Rijksinstituut voor Volksgezondheid en Milieu	RIVM
Netherlands Institute for Health Services Research	Nederlands Instituut voor Onderzoek van de Gezondheidszorg	Nivel
National Childcare Register	Landelijk Register Kinderopvang	LRK
Service Execution Education	Dienst Uitvoering Onderwijs	DUO
Royal Dutch Ice Skaters Association	Koninklijke Nederlandse Schaatsbond	KNSB
Association of Theater and Concert-hall directors	Vereniging van Schouwburg- & Concertgebouwdirecties	VSCD
Association of Dutch Music Venues and Festivals	De Vereniging Nederlandse Poppodia en -Festivals	VNPF
Address Coordinates of the Netherlands	Kadaster	
Association Scientific Research Traffic Safety	Stichting Wetenschappelijk Onderzoek Verkeersveiligheid	SWOV
Dutch Hospital Data	Landelijke Basisregistratie Ziekenhuiszorg	LBZ

Appendix V
GitHub structure

Table 6: GitHub structure and content of folders

File location	Content	Notes
../Buurtvergelijker/Shiny_app	- Server - User interface	
../Buurtvergelijker/Incidents	- Data preparation RWS-dataset - Functions for traffic incidents to create visualizations and filter data	
../Buurtvergelijker/Facilities	- Data preparation CBS-dataset - Functions for amenities dashboard to create visualizations, filter data, postal code look up and top 5 calculation	
../Buurtvergelijker/Data	- Data for amenities dashboard - Part of data for health dashboard (Dekkers, 2022)	- Data on incidents, crime and health need to be downloaded, preprocessed and stored in this folder
../Buurtvergelijker/Crime		Kellij (2022)
../Buurtvergelijker/Health		Dekkers (2022)