# Universiteit Leiden

# Computer Science

**Specialization: Science Based Business**

# Data Visualization Without Privacy Violation


Name: Renuka R. Ramgolam
Date: dd/mm/2018

1st supervisor: Suzan Verberne

2nd supervisor: Arno Knobbe


MASTER'S THESIS


Leiden Institute of Advanced Computer Science (LIACS)

Leiden University

Niels Bohrweg 1

2333 CA Leiden

The Netherlands

# Contents

# List of Tables

# List of figures

# 1. Introduction

Nowadays data is available everywhere. Data of different types, from different sources, regarding different subjects are just one click away. The easy availability of data eases the work and accelerates the speed of working processes. However, not all data is supposed to be easily available or accessible. Amongst all data there are some categories of data that are classified as sensitive data. This data contains sensitive information that may not be public. You can think of personal, financial, legal and health related information. The data is sensitive but still very useful for a responsible person or a group(s) of responsible people.

There are various ways of using data. Data can be used to gain knowledge about something, but it can also be used for planning and analysis, this is called data analytics. Data Analytics is the science of examining raw data with the purpose of finding patterns and drawing conclusions about that information by applying an algorithmic or mechanical process to derive insights [1]. Various business intelligence technologies are developed for data analytics. Business intelligence (BI) is a technology-driven process for analyzing data and presenting actionable information to help executives, managers and other corporate end users make informed business decisions [2]. Data visualization is very commonly used for presenting the results. Data visualization is widely used because it gives quick access to and a clear overview over a huge amount of data. Many companies and governmental institutions are now making a transition to the new way of presenting data through dashboards. These dashboards are used for presenting information to the management and employees. Each end user has a different authorization for certain data. While data visualization is a very beneficial and valued method for presenting information it also contains the risk of giving unauthorized people access to sensitive data.

**Scientific contribution**

The demand for modernization of government agencies is very high in and constantly increasing, but very little research is being done in this specific area. Especially not for Dutch governmental institutions specifically. Combining the computational design process of Ben Fry [46] with the general data storage architecture and additional PPDM techniques can give us a very useful process framework for privacy preserving data visualization.

**Problem statement:** If sensitive data leaks or becomes easily available or accessible it can have disastrous consequences. Sensitive data must be protected and dealt carefully with, without restricting the positive use possibilities for those to whom this information is essential.

**Research question (RQ):** how can sensitive governmental data be visualized and used to full potential in planning without privacy violations?

**Sub-research questions (s-RQ):**

1. What are the necessary steps in the visualization process of citizen data at governmental institutions?
2. Which privacy preserving data mining (PPDM) techniques are useful during the data visualization process of citizen data?

3. Which rules and regulations need to be considered during the data visualization process of citizen data?
4. What are the challenges in visualization of privacy sensitive data if considered the data set size and aggregation level?

Even though the automation and visualization are happening on a broad scale, there are still many confusions, uncertainties and risks that the processes involve. Especially with the new EU General Data Protection Regulation (GDPR) that has been enforced since the 25th of May 2018, processes must be carried out very carefully. The GDPR and the change process of municipalities to be GDPR compliant have been the limiting factors of my research.

## 1.1. Research approach and methods

My research is based on interviews with experts, literature study, longitudinal and cross-sectional survey's and experiments.
For my research I did an internship at an IT consultancy company Motion10, where I worked as a data & analytics consultant. Through Motion10 I could carry out my research at the municipality of Alphen aan den Rijn. At a municipality a lot of personal and sensitive data is stored. This data can be used in various ways for various purposes. The municipality of Alphen aan den Rijn is building a data platform with help from Motion10. This platform is meant for presenting management information. In this data platform project various processes are being automated and various dashboards are being made for different divisions of the municipality. This process contains big amounts of personal and sensitive data. It was therefore an ideal place to do my research. My research was partly parallel and partly linked to the data platform and consisted of two test cases at the Schuldhulpverlening (SHV) department of the municipality. In both cases the aim was to investigate how to visualize the personal and/or sensitive data this department has in a useful way for the users of the responsible division, without privacy violation.

## 1.2. Practical contribution

Governmental agencies have a lot of data available that can be used in so many useful ways for so many advantageous purposes. There are many uncertainties and risks that they are afraid of and therefore hesitate to take this step. Because there is not enough information about how these problems can be solved, they are not likely to take a step towards the unknown.
In this thesis I propose a framework for privacy sensitive data visualization that can be applied to citizen data. With this investigation I want to give more clarity where possible and create new possibilities of using existing data more conveniently. In such a way that the available information is not only informative but goes beyond that. The aim is to perform predictive analysis on the available data. The results of this analysis can then be used to monitor governmental institution's services, decision making and so on.

## 1.3. Privacy

My entire research is done considering the most recent legislation in this area, the General Data Protection Regulation (GDPR) effective May 2018. ***The results of this research may only be used for planning and monitoring their policy and services in such a way that no one's privacy and living comfort is violated and no one feels harassed. The numbers in this report are fake and used only for illustration purposes. They do not represent any reality regarding the municipality of Alphen aan den Rijn and may therefore not be used for any other purposes or in any other analysis or reports except for this one.*** Details regarding my research are described further in this report.

## 1.4. Thesis outline

In the second chapter you can read about the background information that I used. In the third chapter the proposed privacy preserving data visualization framework is described, the fourth chapter is about the methods used for evaluating the framework, the fifth chapter shows the results of my experiments followed by the sixth chapter with the conclusion of my research. And at the end additional information that is excluded from the previous chapters can be found.

# 2. Background

## 2.1. EU General Data Protection Regulation (GDPR)

Data can be categorized in different ways. They can be categorized by type, size, purpose and so on. For this research the focus is on governmental data, personal data and sensitive data. Here are some definitions given to have a better understanding of the information that will be provided further. A data subject is an individual whom the data is about, that individual is subject of the data. A data processor is an entity that processes data on behalf of its customers. The sole responsibility of a data processor is to process the data as per instruction without taking ownership of the data. They only take data as input, process it and generate an output. A data controller however is the entity that determines the purposes, conditions and means of the processing of the data done by the data processor [3][4]. Personal data is data that is related to a living individual that can directly or indirectly be used to identify that person [3]. Sensitive data is defined as personal data that includes racial, ethnic, political, religious, medical, financial, biometric information or information about criminal records or sexual life [3].
The general data protection regulation is a regulation to protect the privacy and data of all citizens of the European Union. GDPR makes its applicability very clear - it will apply to the processing of personal data by controllers and processors in the EU, regardless of whether the processing takes place in the EU or not. The GDPR will also apply to the processing of personal data of data subjects in the EU by a controller or processor not established in the EU, where the activities relate to: offering goods or services to EU citizens (irrespective of whether payment is required) and the monitoring of behavior that takes place within the EU. Non-Eu businesses processing the data of EU citizens will also have to appoint a representative in the EU.
A lot of personal and sensitive information from people is being registered and processed at the municipality. Daily that information is being used in various ways for various reasons. The new legislation is therefore of great importance to them. As the municipality is moving towards a more digital and transparent organization the complexity of keeping control over the data, keeping track of it and protecting it is increasing.

### 2.1.1. What is new in this regulation?

Compared to the previous legislation the GDPR has some significant differences. One of the new requirements is the obligation to maintain a register about the data processing for each case. Every time a client reports at the municipality for a certain problem, a new register must be made were all information regarding that particular case is documented from the beginning till the end. Another difference is the core of the law: accountability. According to the new GDPR the need of registering or processing any type of personal or sensitive data should be well explained and substantiated. Each process should be well documented. In case you are asked to provide reasons why you needed the information, you should immediately be able to proof this with the documents. Reporting any type of data leak is also an obligation now. If a data leak is not reported on time and it gets caught, the responsible organization will endure serious consequences and be charged with a huge fine. In the past there was no such obligation.

For the new privacy legislation more, supervisors are mandatory, and the enforcement instruments have been extended with:

- The Authority for Personal Data (Autoriteit Persoonsgegevens (AP)), the national supervisor;
- The Data protection officer (Functionaris Gegevensbescherming (FG)), appointed by the municipality.

The FG monitors the application and compliance with the GDPR within the organization, ensures that the GDPR is actually executed. He is the regulator, point of contact for citizens, switch between AP and the organization. An FG is expected to have above-average expertise in privacy legislation and the practice of data protection. The FG has an independent position and can give (un) asked for advice on the protection of personal data to the municipal executives. Every municipality is organized differently, some have appointed someone in the house, others appoint an external person. In the past an FG was optional not obliged.
The challenge of privacy protection has increased due to digitization. Personal and sensitive data are digitally elusive. It is very difficult to get a hold on personal and sensitive data, where they are stored, how they are handled, why they are registered, whether it is processed correctly etc. Most teams in local governmental institutions nowadays are managing their own application. In the past all registered information was documented in one place and after some time saved in the archive, but now what is saved in the archive depends on how many people deal with 1 document in their own different ways.
Privacy by design is also a new requirement in the GDPR. Privacy by design implies that the GDPR should be fundamentally processed in the basic framework of your organization. The whole structure and system must be such that data is continuously protected.
As article 25 of the GDPR says:

- Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organizational measures, such as pseudonymization, which are designed to implement data-protection principles, such as data minimization, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects [26].
- The controller shall implement appropriate technical and organizational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed.
  That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility.
- In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons [26].

For my research this article is important, because this should be taken into consideration while designing the framework. By design of the framework already it should not be possible to violate anyone's privacy.

## 2.2. Privacy-preserving data mining techniques

Data mining is a process of extracting relevant information from a large data set. These data sets can have personal or sensitive data that may not be published. Privacy preserving data mining (PPDM) deals with hiding a person's sensitive identity without losing the usability of data. Sensitive identities include private information about persons, companies, and governments that must be suppressed before it is shared or published. As said before, nowadays data is available everywhere and it is being used in various ways for various purposes. This makes privacy-preserving data mining techniques increasingly vital [36]. Especially for governmental institutions moving towards digitization, as they hold the biggest part of personal and sensitive information about its citizens. There are many variants of PPDM techniques. In this section only some of the important and mostly used techniques are described.

### 2.2.1. Pseudonymization

Pseudonymization is a technique that enhances privacy by replacing the most identifying fields within a data record by one or more artificial identifiers, or pseudonyms. There can be a single pseudonym for a collection of replaced fields or a pseudonym per replaced field [21]. Specifically, the GDPR defines pseudonymization in Article 3, as "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information." To pseudonymize a data set, the "additional information" must be "kept separately and subject to technical and organizational measures to ensure non-attribution to an identified or identifiable person" [21]. Pseudonymization does not remove all identifying information from the data but merely reduces the ability to link a dataset with the original identity of an individual [21].

### 2.2.2. Anonymization

Anonymization is a technique that strips data from any information that may serve as an identifier of a data subject. This technique is an irreversible change with or without additional information, with some exceptions if the scrambling method is used [27]. A variety of methods are available. The choice of method depends on the degree of risk and the intended use of the data [21].

- **Directory replacement**
  A directory replacement method involves modifying the name of individuals integrated within the data, while maintaining consistency between values, such as "postcode + city" [21].
- **Scrambling**
  Scrambling techniques involve a mixing or obfuscation of letters. The process can sometimes be reversible.
  For example: Annecy could become Yneanc [21].
- **Masking**
  A masking technique allows a part of the data to be hidden with random characters or other data.
  For example: Pseudonymization with masking of identities or important identifiers. The advantage of masking is the ability to identify data without manipulating actual identities [21].

- **Personalized anonymization**
  This method allows the user to utilize his own anonymization technique. Custom anonymization can be carried out using scripts or an application [21].
- **Blurring**
  Data blurring uses an approximation of data values to render their meaning obsolete and/or render the identification of individuals impossible [21].

### 2.2.3. Randomization

Randomization is an economical and efficient approach for PPDM. This PPDM technique can be described by a sever-client model. It consists of a sever S and n clients C1, C2, ... ,Cn. S needs to collect data from the clients to conduct a certain data mining task. For simplicity, we assume that the data includes only one attribute X. Let xi be the value of X for Ci with $1 \leq i \leq n$. When X is privacy sensitive, the clients may not want to reveal their individual data and thus create a dilemma between data mining and privacy. Randomization can be used to solve this problem. By using randomization each individual data gets perturbated with random noise and the perturbed data is revealed for data mining [28]. Randomization is a technique where your experimental data subjects are chosen arbitrarily from the list. This technique is used to prevent bias. Even though choosing randomly is supposed to be unbiased, there is still chance of hidden biases. To prevent hidden biases some varieties of randomization have been developed [23].

- **Simple random sampling**
  Simple Random Sampling is basically where you draw numbers from a hat, choose a card from a deck or a ball from a bingo machine. You can also assign numbers to participants, or treatments, and use a random number table to choose participants and treatment groups. It's called simple random sample because it's *simple to implement.* However, in practice, it's difficult to use because adequate sampling frames (lists of all possible participants) are sometimes difficult or impossible to find [23].
- **Permuted block randomization**
  Sometimes, just choosing participants randomly isn't enough. You might want to balance your participants into groups, or *blocks*. Permuted block randomization is a way to randomly allocate a participant to a treatment group, while keeping a balance across treatment groups. Each "block" has a specified number of randomly ordered treatment assignments [23].
- **Stratified random sampling**
  Stratified random sampling is used when your target population is split up into strata (characteristics like income level or housing status), and you want to include people from each stratum.
  Once you've defined your strata, you can use simple random sampling to choose elements from within each stratum. How this differs from permuted block randomization is that with PBR, you want to assign people into groups; With Stratified Random Sampling, your participants are already in groups, and you want to evenly sample *from* those groups [23].

- **Covariate adaptive randomization**
  In covariate adaptive randomization, a new participant is sequentially assigned to a particular treatment group by taking into account the specific covariates and previous assignments of participants. Covariate adaptive randomization uses the method of minimization by assessing the imbalance of sample size among several covariates [24].

### 2.2.4. Cloaking

Cloaking is a term used in both emailing and in search engine processing. In emailing it is used to describe the process of hiding the originator of an email. In search engine technology it describes how a Web page exists in two forms: one page is delivered to the user while the other page is deliberately developed so that it is placed at the top of a search engine ranking [29]. The basis of this technique is used in various ways, one is where content is presented to the users based on IP addresses. Two different people with two different IP addresses searching for the same information will see different things. Each one of them will see what they are authorized to see and nothing else.

### 2.2.5. Aggregation

Data aggregation is a type of data and information mining process where data is searched, gathered and presented in a report-based, summarized format to achieve specific business objectives or processes and/or conduct human analysis [33]. This technique has proven to be beneficial for and compliant with the GDPR. By using this technique relevant information can be presented without publishing personal or sensitive information about someone. Generally, people also rather see their data being used this way than any other where their privacy is at risk [31].

### 2.2.6. Quasi-identifiers

Quasi-identifiers are pieces of information that are not unique identifiers, but if combined and correlated correctly can be used to create a unique identifier [30]. The quasi-identifiers can be non-personal and non-sensitive information like place of birth or occupation. But if these two are combined in a specific situation it can lead to a certain living person. This is privacy violation. It is therefore necessary that the quasi-identifiers are also considered while making visualizations. If for example in a visualization you can see how many people above 55years in certain regions have heart problems. This visualization has no unique identifiers, the range of possibilities for ages above 55 is very large and the numbers are aggregated so there is no information on individual level in a region. But it is possible that in small regions with a small amount of people, there is only 1 person with an age above 55. So, whoever lives in the same region and knows the other people in that region will know the person that has heart problems after they have seen the visualization. In this case the non-unique identifiers of age, region and aggregated number of people with a heart problem are combined and used to identify a natural living person.

## 2.3. The data visualization process

The data visualization process consists of 7 steps [44] [45] [46]:

1. Acquire: Obtain the data, whether from a file on a disk or a source over a network.
2. Parse: Provide some structure for the data's meaning and order it into categories.
3. Filter: Remove all but the data of interest.
4. Mine: Apply methods from statistics or data mining to discern patterns or place the data in mathematical context.
5. Represent: Choose a basic visual model, such as a bar graph, list, or tree.
6. Refine: Improve the basic representation to make it clearer and more visually engaging.
7. Interact: Add methods for manipulating the data or controlling what features are visible.

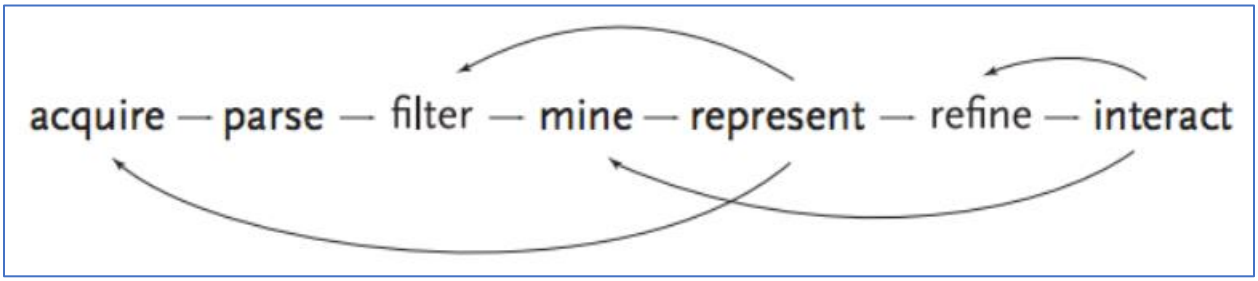



*Figure 1. Computational Design Process [46]*

## 2.4. Data storage architecture

There are various data storage architectures. The **single-tier architecture**, this architecture keeps all the elements of an application, including the interface, middleware and back-end data, in one place. It is used to minimize the amount of data stored and remove redundancy. This architecture is not frequently used in practice [48]. The **two-tier architecture** is a software architecture where the interface and data layer are separated. The data layer gets stored on a server and the interface runs on a client. Due to several limitations like expandability, connectivity and also the limited number of end-users it can support, this architecture is also not often used [48]. The most widely used **three-tier architecture** consists of a bottom, middle and top-tier. The data sources, data warehouse and the application layer are separated [48] [47]. Most companies as well as governmental institutions use these different layers for data storage and usage. Between the tiers Query and ETL tools (Extract, Transform and Load Tools) are used. The query tools can vary between Query and reporting tools, Application Development tools, Data mining tools, OLAP (Online Analytical Processing) tools [48]. During my research the 3-tier data storage infrastructure and the computational design process are used for generating my framework. In this framework the most optimal privacy preserving data mining techniques to be used during the data visualization process are indicated.

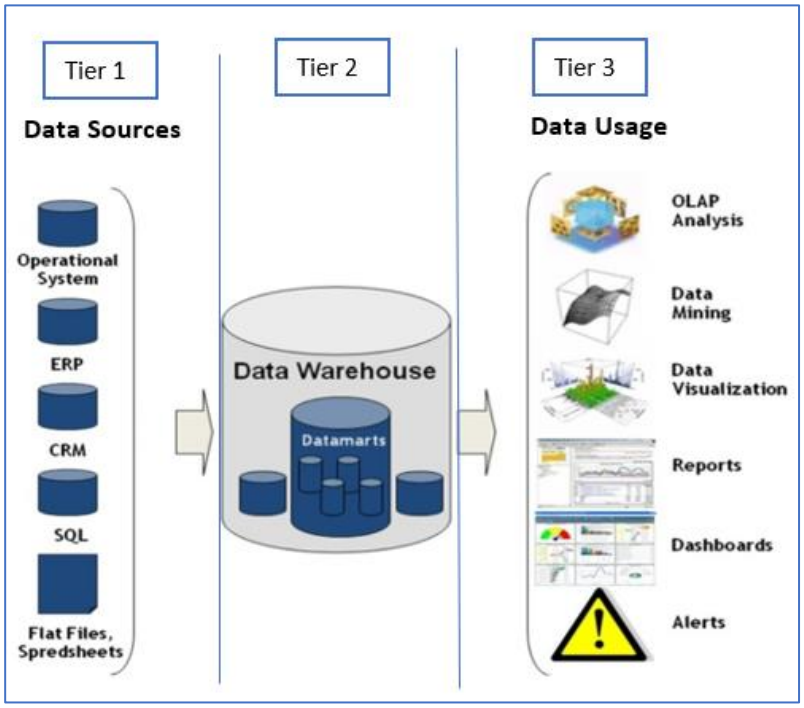



*Figure 2. 3-Tier Data storage architecture*

## 2.5. Data analytics for local governments

Many studies have shown that the world is moving towards digitization [37] [38] [40]. The goal is working towards smart cities, and in this process the government has a very important role. It is also shown by a study in 2013 that while the government has an important role in this process another key issue is the privacy of the citizens [ 37]. According to another research done by the Regulatory Reform and ICT Policy Department of The Ministry of Economic Affairs in July 2016 it is confirmed that the Dutch government is determined to move towards digitization and will make a start for this during the period of 2016 – 2019. Various tasks on various areas helping the Dutch government toward digitization will be executed during this time. They also believe that utilizing data in other ways can be more beneficial for the government, the citizens and the country. In this research one of the most emphasized issues is also the privacy of the citizens that should be protected during this digital transition [40]. In 2017 KPGM did a research and discovered that the use of good quality data analyses at governmental institutions offers the possibility to make relevant predictions about many things. Something that can be used to carry out specific tasks more efficiently and effectively. However, the government also must secure the privacy of the citizens and guarantee the quality of the available data. This makes it appear as though the government is imprisoned in a web of conflicting interests and does not have the potential to fully utilize data [34]. My research is an attempt to make governmental institutions aware and give them more insight about the possibilities they have with the massive amount of data available to them. And show them how they can do that in a way that privacy is preserved while useful and beneficial information is retrieved.

# 3. Framework design

## 3.1. Framework research design

My research is based on qualitative and quantitative research techniques. The qualitative techniques I used are interviews, literature study and surveys. These techniques are used to determine the characteristics of a privacy preserving data visualization framework.

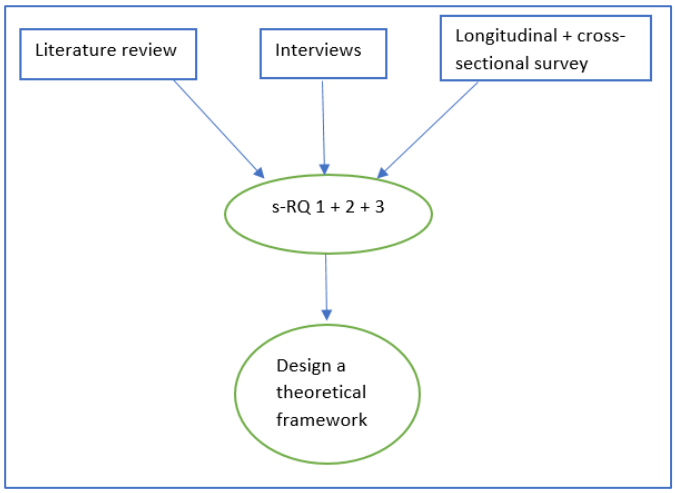



*Figure 3. Framework research design*

Through literature study I have gathered information regarding the process of data visualization, the PPDM techniques and the rules and regulations regarding the data visualization process. To test whether this information is correct, and to gain more insight and information about how it is done practically I have interviewed experts in this field. These experts varied from system administrators, data analysts, privacy officer to consultants. A Longitudinal and Cross-sectional survey is also used to gain more knowledge and insight about the overall process of data visualization. This survey is done at the municipality of Alphen aan den Rijn over a period of 5 months. During these months I gathered information regarding approximately 5 data visualization processes from the beginning till the end. With the gathered information, I defined the characteristics of the data visualization process at local governments and designed a theoretical privacy preserving framework for the data visualization process. In this framework the different steps of the process are described and the PPDM techniques that are used during each step.

## 3.2. Interviews

Interviews can vary from structured, semi-structured to unstructured interviews. I used a semi-structured interview method. I have used a fixed set of questions as a guide to the interview. As these questions and answers to these questions brought forward new ideas during the interview, we continued talking about them. I conducted semi-structured interviews with experts in this field. These experts varied from system administrators, data analysts, privacy officer to consultants. Several interviews were conducted during my research, with different purposes.

These different types of interviews were:

- Interviews with experts in the field of data visualization
  These experts were consultants with advanced experience in data visualization regarding various subjects in various environments (corporate as well as governmental).
- Interviews with representatives from the municipality of Alphen aan den Rijn
  These representatives were members of the data platform project at the municipality.
- Interviews with the privacy officer of the municipality of Alphen aan den Rijn, who was also appointed as privacy officer by the municipality of Amsterdam during my research period.
- Interviews with employees of the department regarding my test cases.

**List of Interviewees**

Expert 1: Senior Data & Analytics Consultant at Motion10
Expert 2: Junior Data & Analytics Consultant at Motion10
Expert 3: Senior Technical Consultant Business Intelligence Motion10
Expert 4: Senior Database Administrator
Expert 5: Data platform project team leader / Financial Advisor for Engineering and City Management at the municipality
Expert 6: Senior Application Manager and Data Analyst
Expert 7: Functional and data manager at the information and automation department / ex senior IT consultant at Centric
Expert 8: Senior System manager
Expert 9: Privacy Officer / Legal Advisor
Expert 10: Advisor & Project manager for information management
Expert 11: Senior manager SHV department

The interviews with these people are taken at different stages multiple times, for different reasons. In the first stage the aim of the interviews was to gather information regarding the data visualization process. During this stage general interviews were conducted with the different groups of people. The same people were also interviewed multiple times during my longitudinal and cross-sectional survey. This survey was also in order to determine the characteristics of the data visualization process. The results of the interviews combined with the information gathered through the literature review and longitudinal and cross-sectional survey was used to define the characteristics of the general data visualization framework. Interviews were also taken in a later stage of the research. These interviews were taken during the test stage of the research. The aim of these interviews was to validate the quality and the privacy preserving aspect of my framework. A more elaborate explanation can be found in Appendix B.

## 3.3. Longitudinal and cross-sectional Survey

Longitudinal research essentially investigates processes across multiple time periods. Since the time duration between data collection efforts is defined by the researcher and by the unit under investigation, the length of a longitudinal study and number of data collection periods vary across designs. Longitudinal designs vary along six parameters: length of study; duration between data collection efforts; number of data collection Periods; method of data collection; research objectives; and unit of analysis [43].

A cross-sectional study is a type of observational study design that involves looking at something that differs on one key characteristic at one specific point in time [49]. During my research I conducted surveys based on the longitudinal survey model, during a period of 5 months. The research started on the 1st of February 2018 and ended on the 1st of July 2018. The subject of this survey was the data visualization process. This survey was held to determine the different stages and characteristics of a data visualization process. The period for each process was 3 weeks. During these 5 months the same process was investigated repeatedly regarding 5 different visualization subjects. The combination with cross-sectional study lies in the fact that for each process the subject was different.

*Table 1. Survey details*

| | |
|---|---|
| **Sponsor/Data owner** | The municipality of Alphen aan den Rijn and Motion10 |
| **Type of study** | Longitudinal + cross-sectional |
| **Subject definition** | Data visualization process using personal and sensitive data |
| **Type of data** | Governmental data |
| **Entire duration** | 5 months |
| **Number of sub-divided processes investigated** | 5 |
| **Duration of each process** | 3 weeks |
| **Subjects of the processes** | Burgerzaken, Vergunningen, Financien, Serviceplein, Openbare Werken |
| **Sample design** | During this research it was necessary that the data models used for the visualization, contained sensitive and/or personal data. |
| **Data collection mode** | Face-to-face interview; Observation |
| **Sample replacement** | During each process the content and therefore the sensitive aspect of the data changed |
| **Data availability** | The data used for this study is stored on servers of the municipality of Alphen aan Den Rijn and is not publicly available. |

Questions asked during the interviews:

- What is the subject of the visualization process?
- What is the sensitive aspect in this visualization process?
- What are the stages in the data visualization process?
- What is done to ensure privacy protection?
- Which privacy preserving data mining (PPDM) techniques are used during the data visualization process?

The list of interviewees can be seen in table 2.

*Table 2. Survey Interviewees*

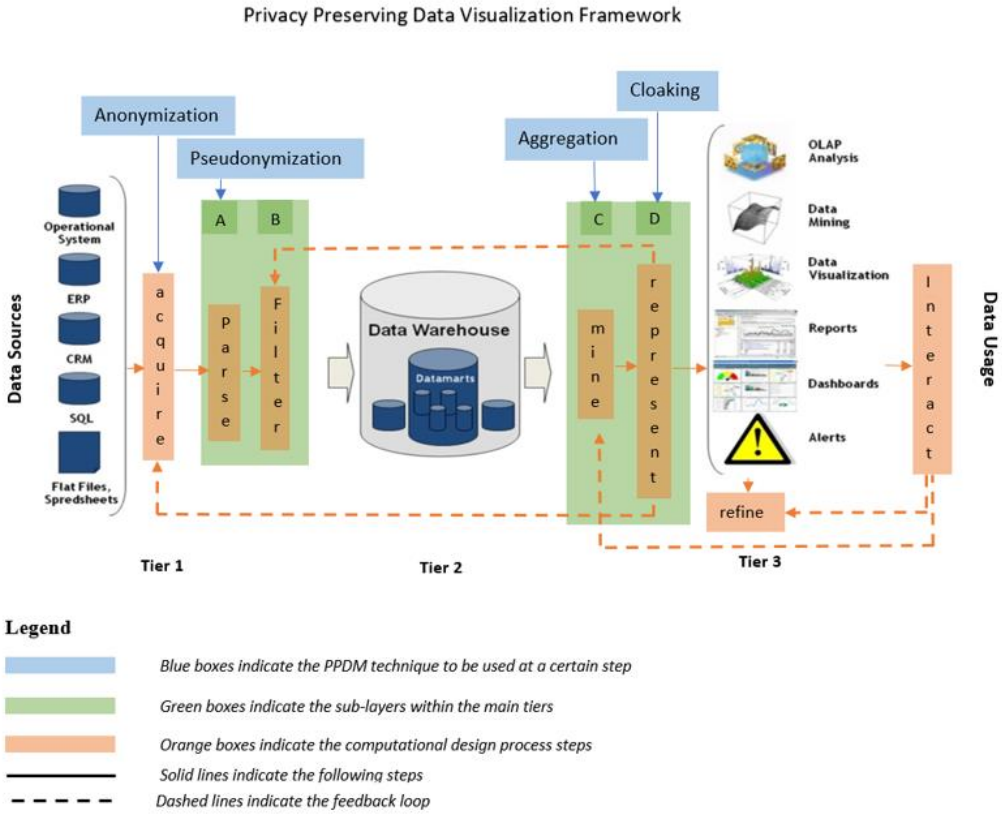
| Group | Interviewees |
|---|---|
| 1 | Expert 1: Senior Data & Analytics Consultant at Motion10<br>Expert 2: Junior Data & Analytics Consultant at Motion10<br>Expert 3: Senior Technical Consultant Business Intelligence Motion10 |
| 2 | Expert 4: Senior Database Administrator<br>Expert 5: Data platform project team leader / Financial Advisor for Engineering and City Management at the municipality<br>Expert 6: Senior Application Manager and Data Analyst<br>Expert 8: Senior System manager |

## 3.4. Characteristics of a privacy preserving data visualization process

After doing literature review and conducting interviews and surveys the following characteristics are determined for a privacy preserving data visualization process:

- 3-tier data storage architecture
- Sub-layers within the tiers (optional)
- PPDM techniques: Anonymization, Pseudonymization, Aggregation and Cloaking
- ETL + Query tools
- Feedback loop for continuous improvement
- The visualization process must be GDPR compliant

## 3.5. Proposed framework for privacy preserving data visualization



# 4. Methods

To validate the privacy preserving aspect of the proposed theoretical privacy preserving data visualization (PPDV) framework, test cases are used. During these tests private/sensitive citizen data is visualized using the PPDV-framework. The used PPDM techniques are full suppression anonymizing technique, generalization anonymization technique, tokenization pseudonymization technique and count aggregation. The re-identification risk and the tradeoff between privacy and data utility, is used to measure the credibility of the proposed PPDV-framework. To define an area were the PPDV-framework will give the most optimal privacy preserving visuals the data set size and aggregation level are used as variables.

## 4.1. PPDM techniques

**Anonymization – suppression**

It replaces tuple or attribute values with special symbol "**" that is instead of the original value. The original value is replaced with some anonymous value throughout the database [52]. During this research this technique is used for anonymizing personal data like name, last name, race, address details (street name and street number), see figure 5. This information was not necessary for the visualization that is why full suppression is used on these columns of data. For implementing full suppression anonymization during the first and second test case a standard framework based on SQL, of the municipality of Alphen aan den Rijn is used. These queries are confidential and may therefore not be put in this thesis. For the experiments of the 3rd test the ARX tool is used to perform full suppression on the necessary columns of the data.

**Anonymization – generalization**

Generalization replaces exact values with a more general description to hide the specific details of people, making the quasi identifiers less identifying. If the value is numeric, it may be changed to a range of values. For example, age attribute value 45 can be changed to range 40 – 60 [52]. This technique is used on the age column of the data sets for ages between 20 and 100 years in intervals of 5 years (see figure 5). For performing this generalization, the ARX tool is used.

**Pseudonymization - tokenization**

Tokenization is a non-mathematical approach that replaces sensitive data with non-sensitive substitutes, referred to as tokens. The tokens have no extrinsic or exploitable meaning or value [53]. Identifiers are replaced with a non-sensitive identifier that traces back to the original data but are not mathematically derived from the original data (i.e. a credit card number is exchanged in a token vault with a randomly generated token number) [54]. This technique is used to represent the people whose information is in the data sets with a client number instead of their names. The client numbers are unique for each person and are randomly generated using a data generating tool Mockaroo.

**Aggregation – count**

The count aggregation method calculates the number of records with a value for the selected attribute [55]. This method is used during the visualizations of the data for gaining more insights. For the visualizations the PowerBI tool is used.

**Cloaking**

In the proposed framework and also during the test cases at the municipality a cloaking based approach is used to give people access to the visualizations. The role level security is based on this technique. Only authorized people can access the visuals with their credentials, other people who try to access it will not be able to sign in or even though they sign in they will see a blank page or only information that they are authorized to see. In this way everyone only sees information they are allowed to see.

| first_name | last_name | age | gender | race | Marital status | country | province | region | Street name | Street number |
|---|---|---|---|---|---|---|---|---|---|---|
| * | * | [20, 25[ | Male | ********** | Married | Netherlands | Provincie Zuid-Holland | Alphen aan den Rijn | ********** | ***** |
| * | * | [20, 25[ | Male | ********** | Registered partnership | Netherlands | Provincie Noord-Brabant | Eindhoven | ********** | ***** |
| * | * | [20, 25[ | Male | ********** | Married | Netherlands | Provincie Gelderland | Nijmegen | ********** | ***** |
| * | * | [25, 30[ | Male | ********** | Registered partnership ended | Netherlands | Provincie Overijssel | Enschede | ********** | ***** |
| * | * | [25, 30[ | Male | ********** | Registered partnership ended | Netherlands | Provincie Zuid-Holland | Delft | ********** | ***** |
| * | * | [25, 30[ | Female | ********** | Registered partnership ended | Netherlands | Provincie Zuid-Holland | Rotterdam postbusnummers | ********** | ***** |
| * | * | [25, 30[ | Male | ********** | Registered partnership ended | Netherlands | Provincie Noord-Brabant | Eindhoven | ********** | ***** |
| * | * | [25, 30[ | Female | ********** | Widow/Widower | Netherlands | Provincie Flevoland | Almere Haven | ********** | ***** |
| * | * | [30, 35[ | Male | ********** | Registered partnership | Netherlands | Provincie Zuid-Holland | Schiedam postbusnummers | ********** | ***** |
| * | * | [30, 35[ | Male | ********** | Married | Netherlands | Provincie Utrecht | Zeist | ********** | ***** |
| * | * | [30, 35[ | Female | ********** | Registered partnership | Netherlands | Provincie Noord-Brabant | Veghel | ********** | ***** |
| * | * | [30, 35[ | Female | ********** | Widow/Widower | Netherlands | Provincie Overijssel | Zwolle | ********** | ***** |
| * | * | [30, 35[ | Female | ********** | Single | Netherlands | Provincie Flevoland | Almere Stad | ********** | ***** |

*Figure 5. Columns with full suppression anonymization technique*

## 4.2. Tools

The basic model of the framework is divided in 3 parts, the data source, the data warehouse and the data usage (application layer). Between these parts you have ETL- and Query tools. For test case 1 and 2 the municipalities data sources, data warehouse and applications and tools were used. But for the third case this basic model had to be imitated by other tools and software's. For the 3rd case the data source was the data generating tool Mockaroo, the warehouse was my hard drive and the data usage layer was my desktop. The ETL- and query tools used are the ARX anonymization tool and Microsoft PowerBI.

## 4.3. Data

During test case 1 and 2 real data from the municipality is used, but for the 3rd case mock data is used. These data sets contain the following information: Client number, first name, last name, age, gender, race, marital status, country, province, region, street name, street number, postal code and debt amount. These categories of information are chosen, because these are the type of sensitive and private citizen data that is registered at governmental institutions. And as my research is based on private and sensitive data of citizen stored at governmental institutions, it is important that the data sets consist of such information. The datasets are generated with a data generating application called Mockaroo (see Appendix C). The assumptions made while generating these datasets are based on the data of the municipality of Alphen aan den Rijn. This is done to generate a dataset that contains realistic information. Table 1 shows how the mock data is related to the data of municipality Alphen aan den Rijn. In the generated data sets the field "country" can be seen as "the municipality", "the provinces" can be seen as "the districts" and "the regions" can be seen as "the areas".

*Table 3. Relationship between mock data and municipality data*

| Data set 3rd case | Data set Alphen a/d Rijn |
|---|---|
| Client number | Client number |
| First name | First name |
| Last name | Last name |
| Age | Age |
| Gender | Gender |
| Race | Race |
| Marital status | Marital status |
| Country | Municipality |
| Province | Districts |
| Region | Areas |
| Street name | Street name |
| Street number | Street number |
| Postal code | Postal code |
| Debt | Debt |

**Client number**

This is a pseudo identification code for the clients, that is used at municipalities. It is a 9digit number that's unique for every client. With Mockaroo a random 9-digit number is generated for every client.

**First name + Last name + Gender + Race + Marital Status**

Random unique male and female names are generated by the tool with according gender information. The race and Marital status are also randomly assigned to the people.

**Age**

The chosen ages for the data set of the third test case is based on the age distribution of people seeking for debt assistance at the municipality of Alphen aan den Rijn. This distribution can be seen in figure 5. The chosen age interval to be used in these data sets is [20,95]. These ages are chosen because as can be seen in figure 6, for ages younger than 20 and higher than 95 the number of debtors are less than 0.5%.

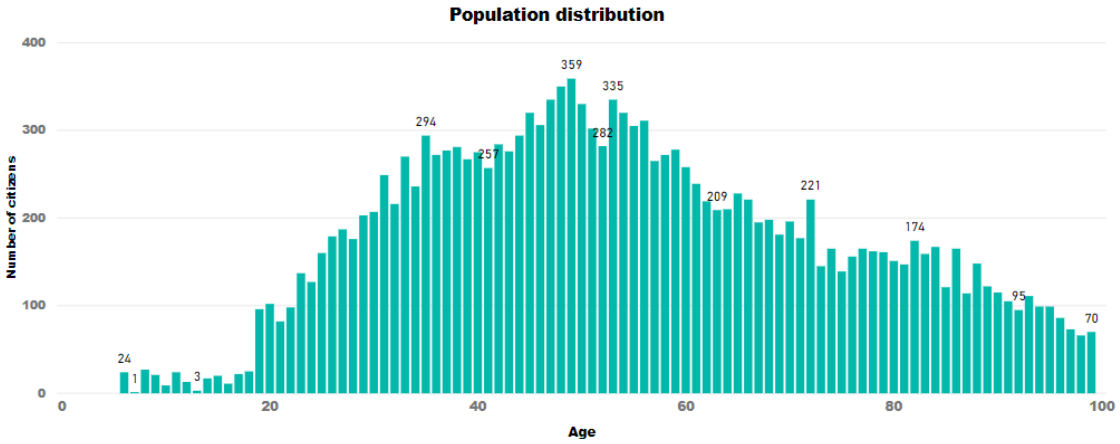



*Figure 6. population distribution of debtors by age at the municipality of Alphen aan den Rijn*

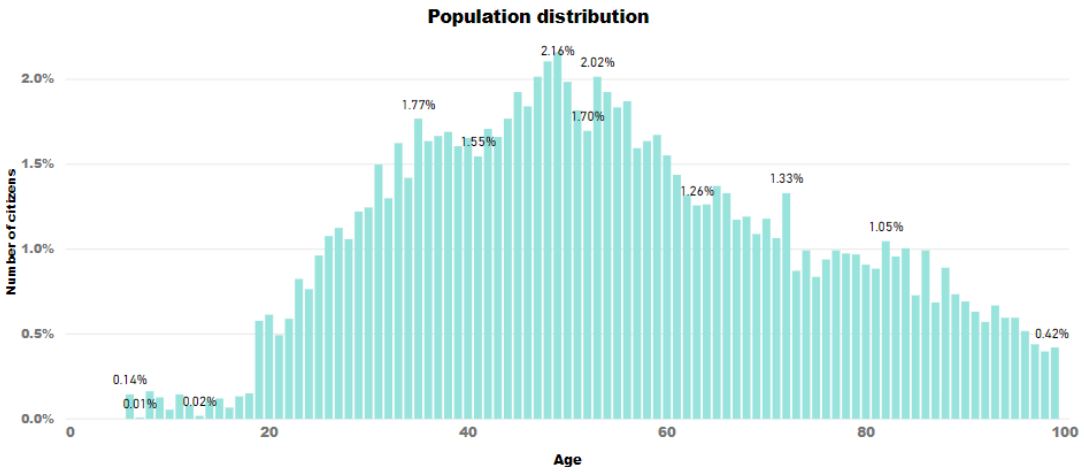



*Figure 7. Percentage display of debtors by age at the municipality of Alphen aan den Rijn*

**Country + Province + Region + Street name + Street number + Postal code**

As country the data sets are restricted to the Netherlands, as this research is mostly for Dutch governmental institutions. The Province, Region and address information are all based on the Netherlands.

**Debt**

This column is a random generated amount of money between €5000,- and €5000000,-. It indicates the amount of the debt someone must pay. This wide range is chosen, because in real scenarios the cases also differ from very small amount of debts to very huge amounts of debts people have to pay and therefore seek help at their municipality.

**Size of data sets**

The population of the Netherlands is approximately 17million. The distribution of the population per municipality can be seen in figure 7. The data for this distribution is taken from: https://en.wikipedia.org/wiki/List_of_municipalities_of_the_Netherlands. This distribution was used to determine the size of the test data sets. From statistics of the municipality of Alphen aan den Rijn approximately 10% of the population is registered for debt assistance. If we now assume that this is the case for every municipality in the Netherlands, we can calculate the approximate number of debtors per municipality. For determining the largest and the smallest data set I used the largest and the smallest municipality of the Netherlands according to the statistics of figure 7. Amsterdam has the largest population with a number of 853000 people and Schiermonnikoog the smallest population with 938 people. From figure 8 we can conclude that according to the assumptions made, most municipalities have a number of debtors below 10000. There are some between 10000 and 20000 and very few municipalities with a very high number. Therefore, the gap between the chosen size of the data sets decreases as they get smaller. The mock data set sizes chosen, based on the distribution of figure 8 can be seen in table 4.

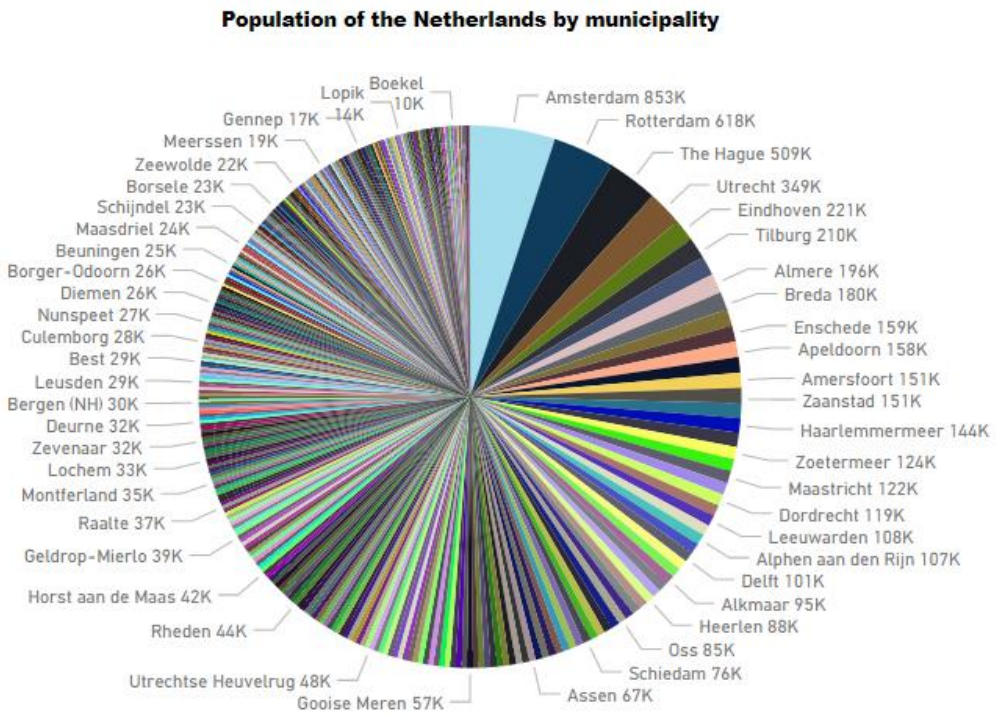



*Figure 8. Population distribution of the Netherlands by municipality*

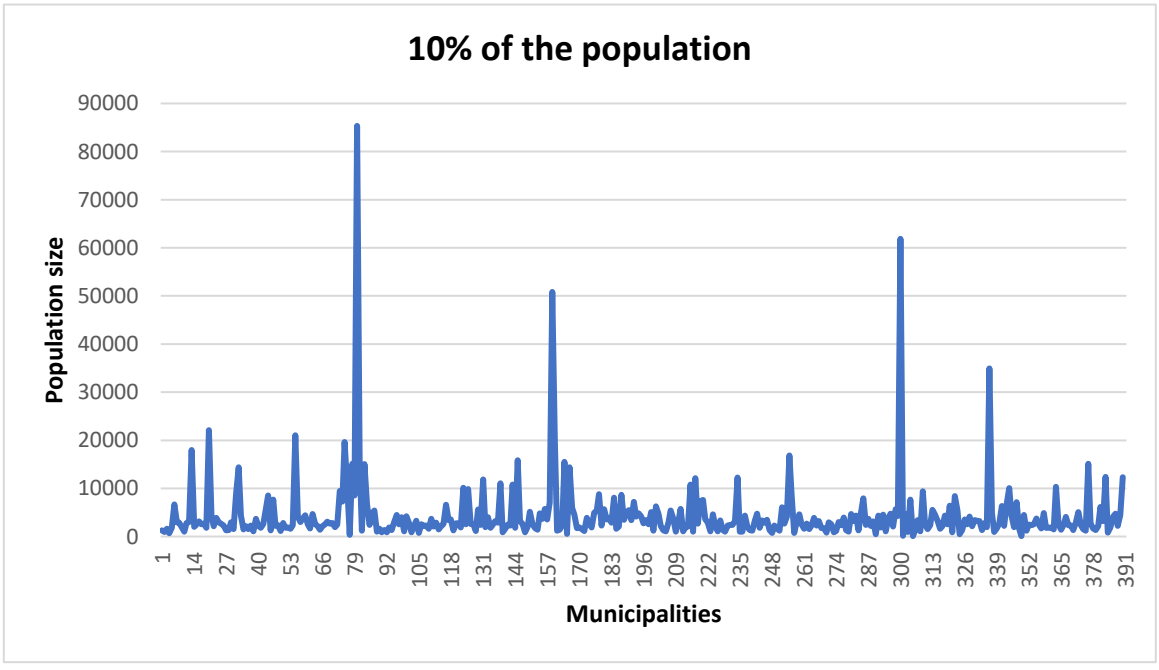



*Figure 9. 10% of the population of Dutch municipalities*

*Table 4. Third test data set sizes*

| Data set | Size |
|---|---|
| **1** | 90000 |
| **2** | 75000 |
| **3** | 60000 |
| **4** | 45000 |
| **5** | 30000 |
| **6** | 20000 |
| **7** | 15000 |
| **8** | 10000 |
| **9** | 9000 |
| **10** | 8000 |
| **11** | 7000 |
| **12** | 6000 |
| **13** | 5000 |
| **14** | 4000 |
| **15** | 3000 |
| **16** | 2000 |
| **17** | 1000 |
| **18** | 500 |
| **19** | 250 |
| **20** | 100 |

## 4.4. Evaluation method

### 4.4.1. Measuring the re-identification risk

Based on these different data set sizes the re-identification risk of anonymized data is measured. This is done to measure what the effect of the data set size is on the privacy preserving aspect of the framework. The ARX tool that calculates this risk based on 3 attacker models, is used to measure this risk. These 3 attacker models of the ARX tool are the prosecutor attacker, the journalist attacker and the marketer attacker model. For re-identifying a specific individual using the *prosecutor re-identification scenario*, the intruder (e.g., a prosecutor) would know that a particular individual (e.g., a defendant) exists in an anonymized database and wishes to find out which record belongs to that individual [50]. While re-identifying an arbitrary individual known as the *journalist re-identification scenario*, the intruder does not care which individual is being re-identified but is only interested in being able to claim that it can be done. In this case the intruder wishes to re-identify a single individual to discredit the organization disclosing the data [50]. The *marketer attacker model* involves re-identifying as many people as possible from the de-identified data even if this means some of them will be incorrectly identified [51]. The results can be seen in table 5 and a graphical visualization of the data in table 5 can be seen in figure 10.

### 4.4.2. Aggregation level

For investigating the impact of aggregation level on the privacy preserving aspect of the visualizations and to define the most optimal range of aggregation level that can be used during the data visualization process using the PPDV-framework, different aggregation levels and

combinations of aggregations are used during the data visualization process of the different sizes of data sets. For each test case different aggregation levels are used. These different aggregation levels per case are also used to determine whether the content of the data set has an impact on the extend of aggregation during the visualizations. The chosen aggregation levels for case 1 are Direct reporting, Toegang, Settlement code. The chosen aggregation levels for case 2 are municipality, region, area codes, gender and marital status. And finally, the chosen aggregation levels for case 3 are Country, Province, Region, Postal code, Age, Gender and Marital Status. These aggregation levels are to investigate at what level or with what combination of aggregations there is a possibility of potential privacy breach. For testing the aggregation levels PowerBI is used. Each anonymized data set is imported into PowerBI and then aggregated on the different levels and combinations of them using the count method. Each aggregation or combination of aggregation is then investigated on possible risky aggregations presented. The results for test case 1,2 and 3 can be seen respectively in table 6,7 and 8. The value of a risky aggregated number is considered to be 1. PowerBI can calculate the number of risky aggregated values in each visualization.

### 4.4.3. Data utility vs. privacy

To test the tradeoff between the privacy preservation and the data utility additional aggregation conditions are applied to the data with a re-identification risk. The conditions to be applied can vary per case depending on the type of data that is visualized and the information the data set contains. An aggregated individual result in a totally privacy preserving visualization can eventually be used to track a natural living person using quasi identifiers. To reduce the re-identification risk per data set on each aggregation level that contains possible risky aggregations, additional changes have to be made to the data set using the appropriate PPDM techniques. The trade off between data utility and privacy is measured using the cluster utility method. This method tests whether the distribution of original data is the same as that of masked data by assigning observations in pooled data to clusters and computing differences between the number of observations from original and masked data for each cluster [56]. In this case the data set with a re-identification risk is compared to the data set where this risk is eliminated with additional aggregation techniques and the difference between the numbers of both visualized data sets gives the change in utility of the data. The results can be seen in table 9 and a graphical visualization of the data in table 9 can be seen in figure 11 and 12.

# 5. Results

In this chapter the results of the investigation regarding the re-identification risk and tradeoff between privacy and data utility of the proposed PPDV-framework based on data set size and aggregation level are presented.

### Re-identification risk

*Table 5. Results re-identification risk test case 3*

| Data set | Size | Re-identification risk (%) | | | Records at risk (%) |
|---|---|---|---|---|---|
| | | Prosecutor | Journalist | Marketer | |
| **1** | **90000** | 3.941 | 3.941 | 3.941 | 1.4 |
| **2** | **75000** | 4.387 | 4.387 | 4.387 | 1 |
| **3** | **60000** | 5.347 | 5.347 | 5.347 | 1.6 |
| **4** | **45000** | 6.891 | 6.891 | 6.891 | 3.35 |
| **5** | **30000** | 9.436 | 9.436 | 9.436 | 7.7 |
| **6** | **20000** | 7.786 | 7.786 | 7.786 | 4.4 |
| **7** | **15000** | 9.711 | 9.711 | 9.711 | 8.8 |
| **8** | **10000** | 7.896 | 7.896 | 7.896 | 5.7 |
| **9** | **9000** | 8.387 | 8.387 | 8.387 | 5.7 |
| **10** | **8000** | 9.173 | 9.173 | 9.173 | 6.8 |
| **11** | **7000** | 10.156 | 10.156 | 10.156 | 10 |
| **12** | **6000** | 11.276 | 11.276 | 11.276 | 13.1 |
| **13** | **5000** | 12.459 | 12.459 | 12.459 | 15.9 |
| **14** | **4000** | 9.466 | 9.466 | 9.466 | 7 |
| **15** | **3000** | 9.633 | 9.633 | 9.633 | 7.9 |
| **16** | **2000** | 9.521 | 9.521 | 9.521 | 7.9 |
| **17** | **1000** | 14.183 | 14.183 | 14.183 | 10 |
| **18** | **500** | 22.666 | 22.666 | 22.666 | 55 |
| **19** | **250** | 8.130 | 8.130 | 8.130 | 4.4 |
| **20** | **100** | 15 | 15 | 15 | 3 |
| **Average** | | **9.772** | | | **9.03** |

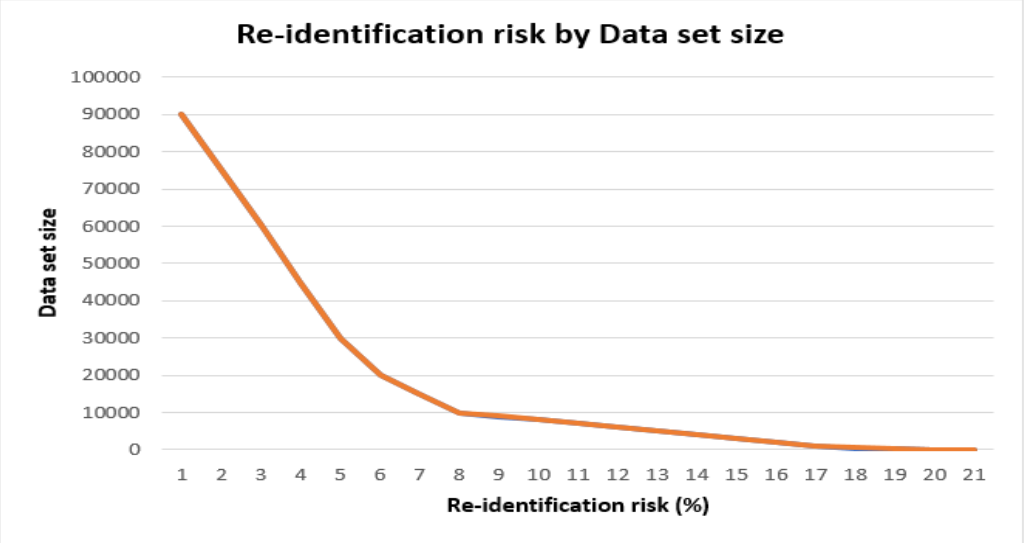



*Figure 10. Re-identification risk by data set size*

As mentioned in the previous section, the re-identification risk is used as a measure to investigate the impact of the data set size on the privacy aspect of the proposed PPDV-framework. Table 5 presents the results of this investigation and figure 10 is a graphical representation of the data in table 5. If we look at the data in table 5 and the graph in figure 10, we can clearly see that the re-identification risk increases as the data set size decreases. This means that data visualization of smaller data sets with private/sensitive data has a higher risk of privacy breach than large data sets with private/sensitive data. According to the results the average chance of re-identification for the PPDV-framework is approximately 9%.

### Aggregation level

*Table 6. Results of different aggregation levels test case 1*

| | Aggregation level | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Data set size | Toegang | Direct reporting | Settlement code |
| 700 | V | V | V |
| 311 | V | V | V |

*Table 7. Results of different aggregation levels test case 2*

| | Aggregation level | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Data set size | municipality | region | Area codes | gender | Marital status |
| ±21000 | V | X | X | X | X |

*Table 8. Results of different aggregation levels test case 3*

| | Aggregation level combination | | | | | | |
|---|---|---|---|---|---|---|---|
| Data set size | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 90000 | V | V | V | V | V | X | X |
| 75000 | V | V | V | V | X | X | X |
| 60000 | V | V | V | V | X | X | X |
| 45000 | V | V | V | V | X | X | X |
| 30000 | V | V | V | V | X | X | X |
| 20000 | V | V | V | V | X | X | X |
| 15000 | V | V | V | V | X | X | X |
| 10000 | V | V | V | V | X | X | X |
| 9000 | V | V | V | V | X | X | X |
| 8000 | V | V | V | V | X | X | X |
| 7000 | V | V | V | V | X | X | X |
| 6000 | V | V | V | V | X | X | X |
| 5000 | V | V | V | V | X | X | X |
| 4000 | V | V | V | V | X | X | X |
| 3000 | V | V | V | V | X | X | X |
| 2000 | V | V | V | V | X | X | X |
| 1000 | V | V | V | X | X | X | X |
| 500 | V | V | X | X | X | X | X |
| 250 | V | V | X | X | X | X | X |
| 100 | V | V | X | X | X | X | X |

| Legend |
|---|
| V – No possible privacy violating aggregated numbers presented if used independently or in combination with previous aggregation levels |
| X – Possible Privacy violating aggregated numbers presented if used in combination with previous aggregation level (but not if used independently for data set sizes 90000 – 2000) |

As mentioned in the previous section for investigating the impact of aggregation level on the privacy preserving aspect of the visualizations and to define the most optimal range of aggregation level that can be used during the data visualization process using the PPDV-framework, different aggregation levels and combinations of aggregations are used during the data visualization process of the different sizes of data sets. For each test case different aggregation levels are used. These different aggregation levels per case are also used to determine whether the content of the data set has an impact on the extend of aggregation during the visualizations. Test case 1 has been done at the municipality of Alphen aan den Rijn. During this case the subject of the data visualization process was citizens of Alphen aan den Rijn with a debt, seeking for help at the municipality. The aim of this process was to visualize how many citizens are registered at the municipality for debt assistance and through which way they registered themselves, as there are 2 different ways to come to the debt assistance department. In table 6 the results of the data visualization process of the first test case can be seen. Test case 2 has also been done at the municipality of Alphen aan den Rijn. During this case the subject of the data visualization was the geographical division of the citizens of Alphen aan den Rijn with a debt further categorized by marital status and gender. This test case contained more private and sensitive data than the previous one. The results for test case 2 can be seen in table 7. The subject of the 3rd case was visualization of the people with a debt in the Netherlands. During this case not only, the aggregation levels are varied, but the data set size as well. Table 8 presents the results of test case 3. If we consider the results of table 8 we can see that for large data sets there is no privacy threat on more detailed aggregation levels compared to smaller data sets. The same goes for combining the different aggregation levels. For the largest data set the privacy threat occurs when aggregation level 1 to 5 are used in combination with aggregation level 6 and/or 7. For the data sets between 75000 and 2000 the combination of aggregation level 1, 2, 3 and 4 give no risky aggregations. But when these 4 together are combined with 5, 6 and/or 7, there are possible privacy threats in the visualizations. For the data set of size 1000 the threats start on aggregation level 3 and for the data sets smaller than 1000 the threats start at aggregation level 2 already. But if we compare the results of table 6 with the results of table 8 for the same data set size we can see that the privacy threat is also dependent on the information in the data set. A small data set with less private/sensitive data compared to a data set with more private/sensitive data of the same size can have a huge difference in possible privacy breach on the same aggregation levels.

**Data utility vs. privacy**

As mentioned in the previous chapter the tradeoff between the data utility and privacy preservation of the visualizations are measured using the cluster utility method. The results of the experiments done regarding this subject can be seen in table 9. Figure 11 shows a graphical representation of the change in data utility when privacy preservation is increased per aggregation level and figure 12 shows a graphical representation of the change in data utility when privacy preservation is increased per data set for each different size.

*Table 9. Data utility vs. Privacy*

| Change in data utility if privacy preservation is increased (%) | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Aggregation level | | | | | | |
| **Data set size** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **90000** | 0 | 0 | 0 | 0 | 0 | 0.05 | 5.7 |
| **75000** | 0 | 0 | 0 | 0 | 0.05 | 0.23 | 8.4 |
| **60000** | 0 | 0 | 0 | 0 | 0.06 | 0.3 | 13.4 |
| **45000** | 0 | 0 | 0 | 0 | 0.1 | 0.46 | 21.9 |
| **30000** | 0 | 0 | 0 | 0 | 0.17 | 0.73 | 36.8 |
| **20000** | 0 | 0 | 0 | 0 | 0.25 | 0.64 | 28 |
| **15000** | 0 | 0 | 0 | 0 | 0.2 | 1.1 | 38.5 |
| **10000** | 0 | 0 | 0 | 0 | 0.64 | 0.9 | 27.31 |
| **9000** | 0 | 0 | 0 | 0 | 0.73 | 1.3 | 32.3 |
| **8000** | 0 | 0 | 0 | 0 | 0.78 | 1.68 | 37.16 |
| **7000** | 0 | 0 | 0 | 0 | 0.78 | 1.92 | 42.65 |
| **6000** | 0 | 0 | 0 | 0 | 0.7 | 3 | 47.8 |
| **5000** | 0 | 0 | 0 | 0 | 0.86 | 4.62 | 55.18 |
| **4000** | 0 | 0 | 0 | 0 | 1.2 | 1.9 | 38 |
| **3000** | 0 | 0 | 0 | 0 | 29.2 | 54.5 | 89 |
| **2000** | 0 | 0 | 0 | 0 | 1.25 | 1.55 | 36 |
| **1000** | 0 | 0 | 0 | 0.2 | 65.2 | 81.2 | 97 |
| **500** | 0 | 0 | 2 | 4 | 64 | 77.8 | 96.8 |
| **250** | 0 | 0 | 7.6 | 17.6 | 20 | 47.2 | 87.2 |
| **100** | 0 | 0 | 34 | 57 | 94 | 94 | 98 |
| **Average** | **0** | **0** | **2.18** | **3.94** | **14** | **18.7** | **46.8** |

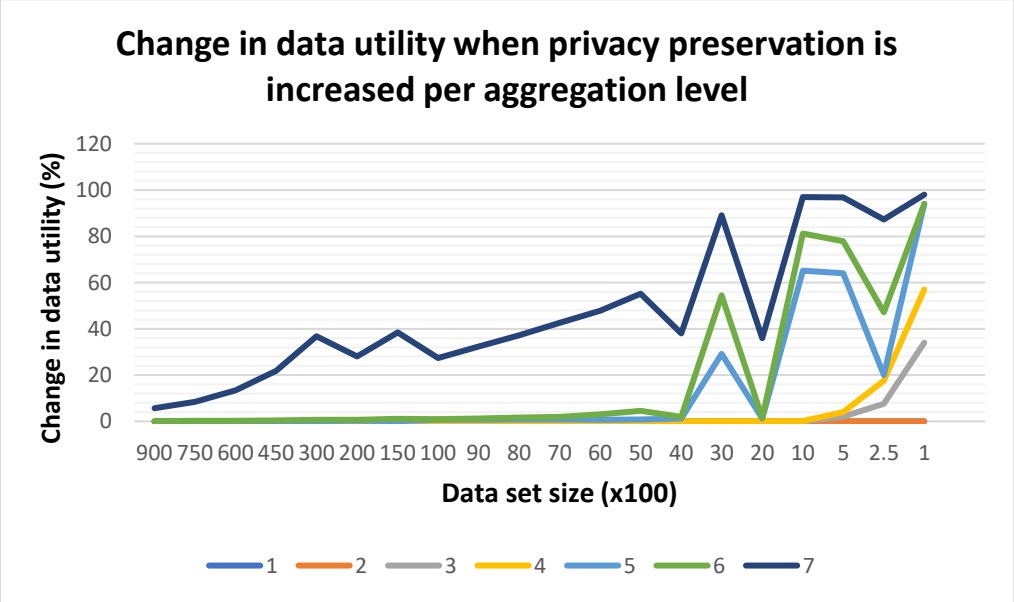



*Figure 11. Change in data utility vs. privacy per aggregation level*

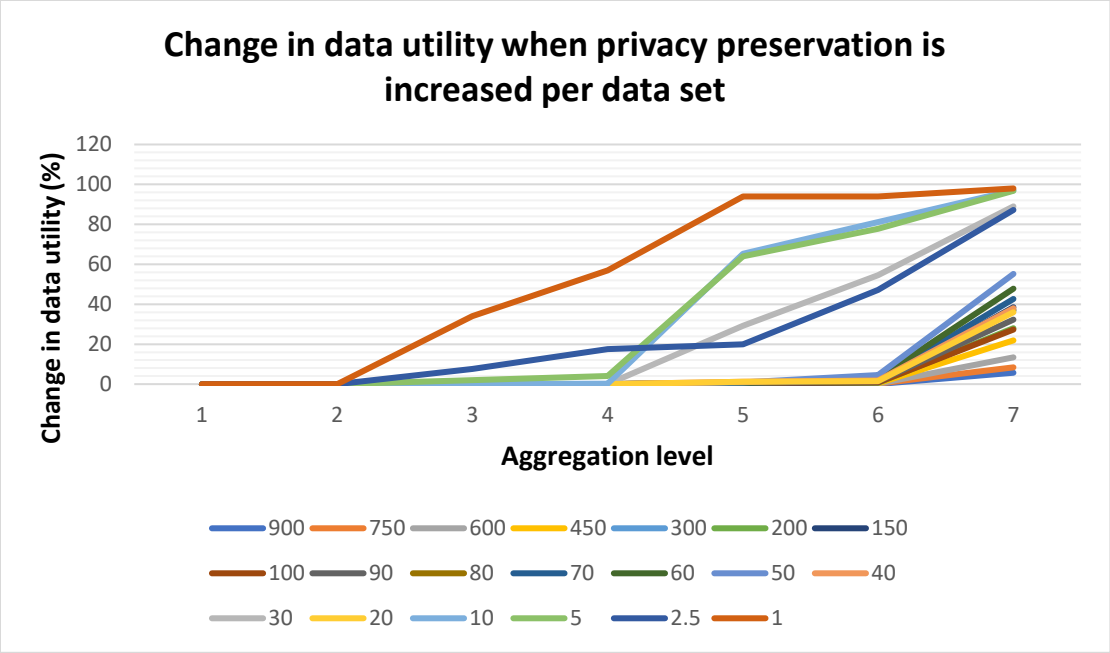



*Figure 12. Change in data utility vs. privacy per data set*

If we consider the results in table 9, figure 11 and 12, we can see that the accuracy of data utility is high more general aggregation levels. As soon as the aggregation levels become more specific and more detailed the accuracy of data utility decreases if the privacy preservation is kept high. If we consider the change in accuracy of the data utility based on the data set sizes, we see that the larger data sets have smaller changes compared to the smaller ones. If we look at the graph in figure 11 we can see that the aggregation levels 1,2,3 and 4 are the aggregation levels with the minimum change in data utility. The average change in data utility for these four aggregation levels is approximately 3%. If we look at the graph in figure 12 we can see that the data sets with sizes between 90000 to 2000 have a minimum change in data utility for aggregation levels 1 to 6.

# 6. Conclusion

The aim of this research was to investigate how sensitive governmental data can be visualized and used to full potential in planning without privacy violations? After completing this research, it can be concluded that governmental institutions can use private/sensitive data in visualization for planning and monitoring by using the proposed privacy preserving data visualization framework (PPDV-framework) in this thesis. In order to find a solution to the main research question the research was split into four sub-research subjects. The first sub-research focused on what the necessary steps in the visualization process of citizen data at governmental institutions are? Research showed that, it is necessary that the data to be visualized is digitally available, the institution has a 3-tier data storage architecture, reliable ETL + query tools and that the correct anonymization, pseudonymization, aggregation and cloaking techniques are used during the visualization process. This last point lead to the second sub-research subject: which privacy preserving data mining (PPDM) techniques are useful during the data visualization process of citizen data? It can be concluded that the following PPDM techniques should be used during the visualization process: full suppression anonymization for data that is not necessary for the visualization, generalization anonymization and the count aggregation function for private/sensitive data that is necessary for the visualization. A cloaking based technique must be used for accessing the visualizations. The third sub-research subject focused on which rules and regulations need to be considered during the data visualization process of citizen data? Since May 2018 the new general data protection regulation has to be used for everything regarding private/sensitive data. Therefore, each visualization process must also be GDPR compliant, as private/sensitive data is used for the visualizations. This means each visualization process must be well documented and registered. The need of processing any type of personal or sensitive data should be well explained and substantiated. Privacy by design is also a requirement that should be considered during each process and implies that the GDPR should be fundamentally processed in the basic framework of the visualization process. The last sub-research subject focused on what the challenges are in visualization of privacy sensitive data if considered the data set size and aggregation level? Based on the results of the experiments it has been proved that the data set size and the aggregation level can be challenging in privacy preserving data visualization processes. Both variables have an impact on the privacy preservation and accuracy of the data utility. Based on the results it can be concluded that visualization of data sets larger than 2000, where aggregation levels 1 to 4 are used give the most privacy preserving and accurate results. For data sets smaller than 1000, the risk of privacy violation increases. How big the risk is depends on the amount of private/sensitive data the data set contains.

In short, it can be concluded that governmental institutions must define clearly what information is necessary for their insights and to what aggregation level they need to go for having the right information from the visualizations. If the parameters are well chosen and used in the PPDV-framework combined with the proposed PPDM techniques, governmental institutions can get more valuable information from the data they already have.

# References


1. Avantika Monnappa (February 7, 2018): Data Science vs. Big Data vs. Data Analytics https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article
2. Margaret Rouse ( August 2017): B**usiness intelligence (BI)** http://searchbusinessanalytics.techtarget.com/definition/business-intelligence-BI
3. Eric Tierling (May 2017): How Microsoft Azure Can Help Organizations Become Compliant with the EU General Data Protection Regulation (GDPR)
4. Techopedia: Data processor https://www.techopedia.com/definition/18977/data-processor
5. Paul Voigt; Axel von dem Bussche (2017): The EU General Data Protection Regulation (GDPR)
6. Tal Z. Zarsky (2011): Governmental Data Mining and its Alternatives Provided by: Available through Universiteit Leiden / LUMC
7. Milda Macenaite (2017): From universal towards child-specific protection of the right to privacy online: Dilemmas in the EU General Data Protection Regulation
8. Claudio Bettini; Sushil Jajodia; Pierangela Samarati; X. Sean Wang (June 2009): Privacy in Location-Based Applications
9. Cynthia Dwork (January 2011): A Firm Foundation for Private Data Analysis
10. M. Prakash; G. Singaravel (25 February 2015): An approach for prevention of privacy breach and information leakage in sensitive data mining
11. Susanne Bahn; Pamela Weatherill (2012): Qualitative social research: a risky business when it comes to collecting 'sensitive' data
12. Smith, V. S. (2013). Data dashboard as evaluation and research communication tool. In T. Azzam & S. Evergreen (Eds.), Data visualization, part 2. New Directions for Evaluation, 140, 21–45.
13. Eric Tierling (May 2017): How Microsoft Azure Can Help Organizations Become Compliant with the EU General Data Protection Regulation (GDPR)
14. Alexandre Evfimievski; Johannes Gehrke; Ramakrishnan Srikant (June 2003): Limiting Privacy Breaches in Privacy Preserving Data Mining
15. Geoff Webb (18 Augustus 2005): A Framework for Evaluating Privacy Preserving Data Mining Algorithms
16. José Ramón Padilla-López; Alexandros Andre Chaaraoui; Francisco Flórez-Revuelta (2015): Visual Privacy protection methods: A Survey
17. Michele Drgon, Gail Magnuson, and John Sabo (17 May 2016): Privacy Management Reference Model and Methodology (PMRM) Version 1.0.
18. George Danezis; Josep Domingo-Ferrer; Marit Hansen; Jaap-Henk Hoepman; Daniel Le Métayer; Rodica Tirtea; Stefan Schiffner (December 2014): Privacy and Data Protection by Design – from policy to engineering
19. Ali Khoshgozaran; Cyrus Shahabi: Private Information Retrieval Techniques for Enabling Location Privacy in Location-Based Services
20. Clyde Williamson (January 5, 2017): Pseudonymization Vs. Anonymization And How They Help With GDPR
21. GDPR Report (28th September 2018): Data masking: anonymization or pseudonymization?

22. Matt Wes (25 April 2017): Looking to comply with GDPR? Here's a primer on anonymization and pseudonymization
23. Stephanie (June 27, 2016): Randomization in Statistics and Experimental Design
24. KP Suresh (2011): An overview of randomization techniques: An unbiased assessment of outcome in clinical research
25. Dr. Urmila Aswar (December 30, 2013): Randomization techniques
26. Intersoft Consulting; Art. 25 GDPR Data protection by design and by default; Available at: https://gdpr-info.eu/art-25-gdpr/
27. Waltraut Kotschy: The new General Data Protection Regulation – Is there sufficient pay-off for taking the trouble to anonymize or pseudonymize data?
28. Yu Zhu; Lei Liu (2004): Optimal Randomization for Privacy Preserving Data Mining
29. Darrel Ince (2013): A Dictionary of the Internet (3 ed.)
30. The Organization of Economic Co-operation and Development (2007): Glossary of Statistical Terms
31. Danvers Baillieu (March 9, 2018): Why aggregated data benefits everyone in a GDPR world
32. Margaret Rouse: Data aggregation; Available at: https://searchsqlserver.techtarget.com/definition/data-aggregation
33. Techopedia: Data aggregation; Available at: https://www.techopedia.com/definition/14647/data-aggregation
34. E.N. Berfelo LLM; A.P.H. de Klerk MSc; J.W. Eikema LLM; M.H.N.M. Hermans BSc (February 2017): Data Analysis in the Public Sector
35. Criterion Content Team (April 26, 2016): How data analytics can improve local government
36. Vinoth kumar J.; Santhi V. (2016): A Brief Survey on Privacy Preserving Techniques in Data Mining
37. Antoni Martínez-Ballesté; Pablo A. Pérez-Martínez; Agusti Solanas (June 2013): The Pursuit of Citizens' Privacy: A Privacy-Aware Smart City Is Possible
38. Rida Khatoun; Sherali Zeadally (August 2016): 'Smart Cities: Concepts, Architectures, Research Opportunities', Communications of the ACM: Vol. 59 No. 8, Pages 46-57
39. Microsoft SQL Server (Februari 2017): Het zevenlagenmodel en de opbouw van een Flexibel Microsoft dataplatform
40. The Ministry of Economic Affairs; Regulatory Reform and ICT Policy Department (July 2016): Digital Agenda for the Netherlands
41. Emanuelle Alm; Niclas Colliander; Fredrik Lind; Ville Stohne; Olof Sundstrom; Maikel Wilms; Marty Smits (June 2016): Digitizing the Netherlands
42. Dale Edgar (2000): Data Sanitazion Techniques
43. Alladi Venkatesh; Nicholas P. Vitalari (1991): Longitudinal Surveys in Information Systems Research: An Examination of Issues, Methods, and Applications
44. Ben Fry: Visualizing Data; Chapter 1. The Seven Stages of Visualizaing Data; Available at: https://www.oreilly.com/library/view/visualizing-data/9780596514556/ch01.html

45. Quora; What are the required steps to create a successful data visualization? ; Available at: https://www.quora.com/What-are-the-required-steps-to-create-a-successful-data-visualization
46. Benjamin Jotham Fry (2004): Computational Information Design
47. Tutorials Point (2014): Data Warehousing
48. Guru99; Data Warehouse Concepts, Architecture and Components ; Available at: https://www.guru99.com/data-warehouse-architecture.html
49. Maninder Singh Setia (2016 May-Jun): Methodology Series Module 3: Cross-sectional Studies
50. Khaled El Emam; Fida Kamal Dankar (2008): Protecting Privacy Using k-Anonymity
51. Privacy Analytics (November 2015): Re-identification attacks
52. Disha Dubli; D.K Yadav (May-June 2017): Secure Techniques of Data Anonymization for Privacy Preservation
53. Alex Ewerlöf (May 2018): GDPR pseudonymization techniques
54. Anna Pouliou (November 2017): Pseudonymization and De-identification Techniques
55. Microsoft ( 29-08-2018): Aggregaties in Power BI-visualisaties
56. Karr Alan; Oganian Anna; J.P. Reiter; Woo Mi-Ja. (2018): New Measures of Data Utility.

# Appendix A

Complete overview of first dashboard visualizations:

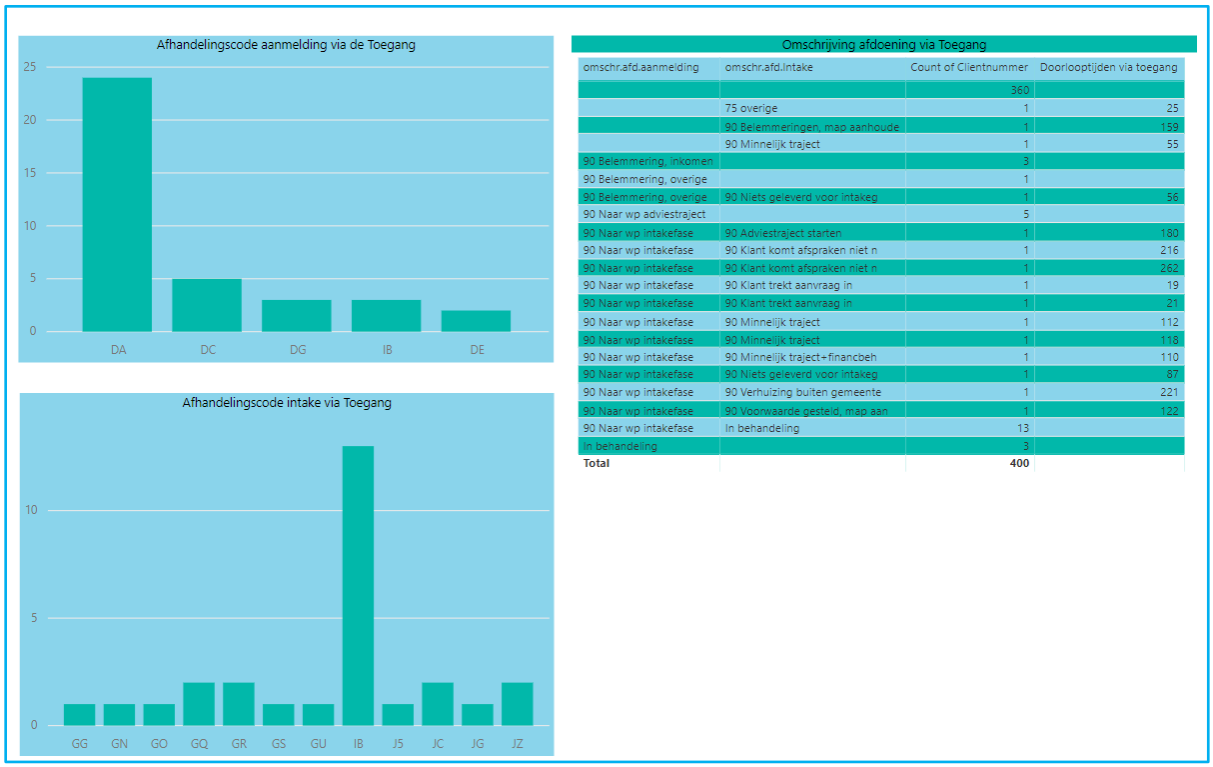
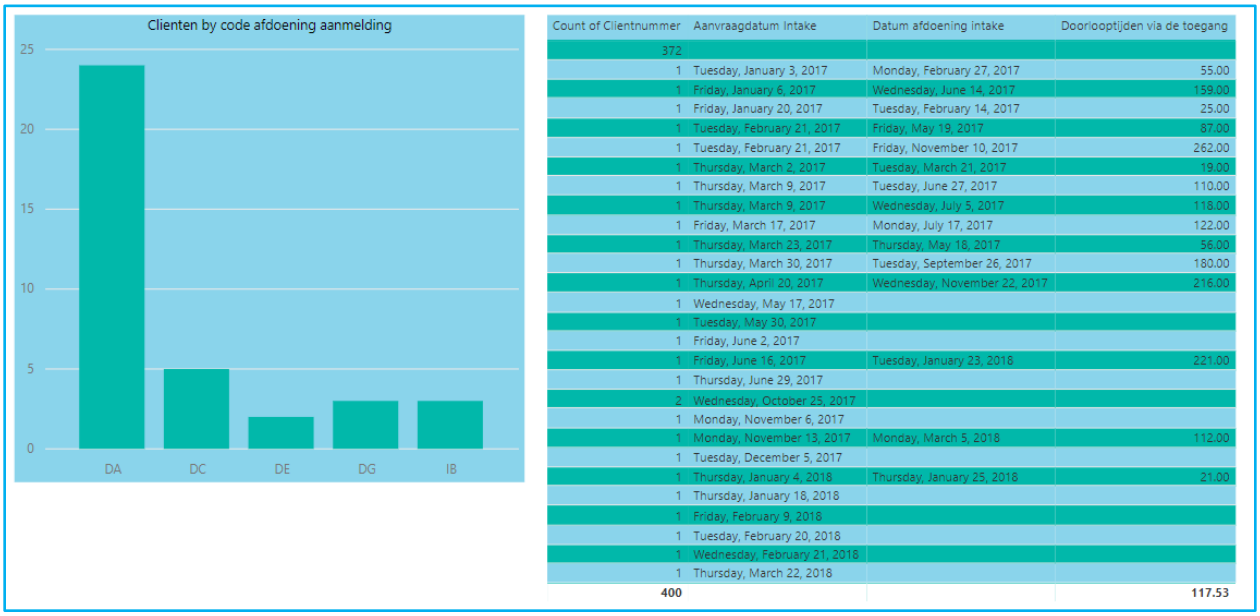


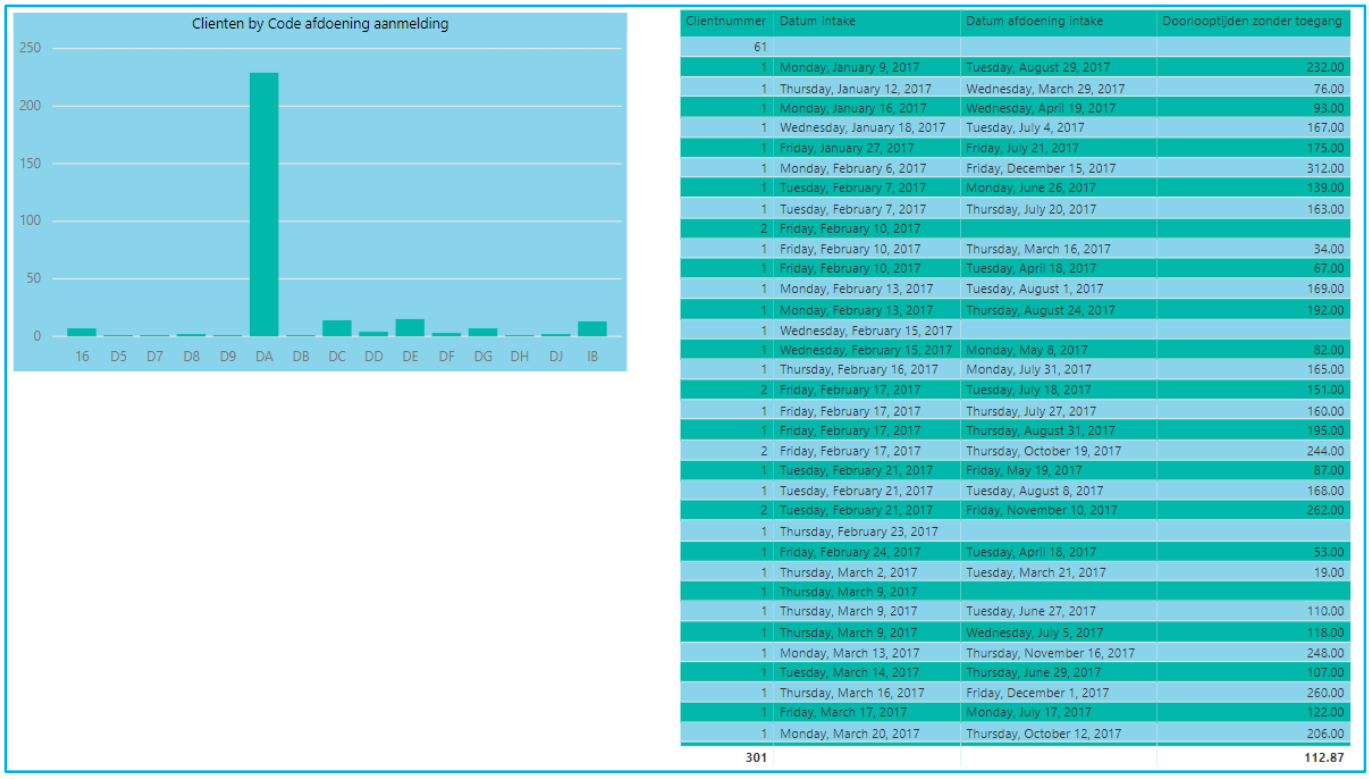

*Figure 13 First tab of the dashboard with clients coming through the "toegang" divided by decision codes in phase 1 and 2*


*Figure 14 Second tab of the dashboard with clients coming through direct reporting divided by decision codes in phase 1 and 2*




| Count of Clientnummer | Aanvraagdatum intake | Datum afdoening intake | Doorlooptijden via de toegang |
|---|---|---|---|
| 372 | | | |
| 1 | Tuesday, January 3, 2017 | Monday, February 27, 2017 | 55.00 |
| 1 | Friday, January 6, 2017 | Wednesday, June 14, 2017 | 159.00 |
| 1 | Friday, January 20, 2017 | Tuesday, February 14, 2017 | 25.00 |
| 1 | Tuesday, February 21, 2017 | Friday, May 19, 2017 | 87.00 |
| 1 | Tuesday, February 21, 2017 | Friday, November 10, 2017 | 262.00 |
| 1 | Thursday, March 2, 2017 | Tuesday, March 21, 2017 | 19.00 |
| 1 | Thursday, March 9, 2017 | Tuesday, June 27, 2017 | 110.00 |
| 1 | Thursday, March 9, 2017 | Wednesday, July 5, 2017 | 118.00 |
| 1 | Friday, March 17, 2017 | Monday, July 17, 2017 | 122.00 |
| 1 | Thursday, March 23, 2017 | Thursday, May 18, 2017 | 56.00 |
| 1 | Thursday, March 30, 2017 | Tuesday, September 26, 2017 | 180.00 |
| 1 | Thursday, April 20, 2017 | Wednesday, November 22, 2017 | 216.00 |
| 1 | Wednesday, May 17, 2017 | | |
| 1 | Tuesday, May 30, 2017 | | |
| 1 | Friday, June 2, 2017 | | |
| 1 | Friday, June 16, 2017 | Tuesday, January 23, 2018 | 221.00 |
| 1 | Thursday, June 29, 2017 | | |
| 2 | Wednesday, October 25, 2017 | | |
| 1 | Monday, November 6, 2017 | | |
| 1 | Monday, November 13, 2017 | Monday, March 5, 2018 | 112.00 |
| 1 | Tuesday, December 5, 2017 | | |
| 1 | Thursday, January 4, 2018 | Thursday, January 25, 2018 | 21.00 |
| 1 | Thursday, January 18, 2018 | | |
| 1 | Friday, February 9, 2018 | | |
| 1 | Tuesday, February 20, 2018 | | |
| 1 | Wednesday, February 21, 2018 | | |
| 1 | Thursday, March 22, 2018 | | |
| 400 | | | 117.53 |

*Figure 15 Third tab of the dashboard with average lead times per category for clients coming through the "toegang"*




| Clientnummer | Datum intake | Datum afdoening intake | Doorlooptijden zonder toegang |
|---|---|---|---|
| 61 | | | |
| 1 | Monday, January 9, 2017 | Tuesday, August 29, 2017 | 232.00 |
| 1 | Thursday, January 12, 2017 | Wednesday, March 29, 2017 | 76.00 |
| 1 | Monday, January 16, 2017 | Wednesday, April 19, 2017 | 93.00 |
| 1 | Wednesday, January 18, 2017 | Tuesday, July 4, 2017 | 167.00 |
| 1 | Friday, January 27, 2017 | Friday, July 21, 2017 | 175.00 |
| 1 | Monday, February 6, 2017 | Friday, December 15, 2017 | 312.00 |
| 1 | Tuesday, February 7, 2017 | Monday, June 26, 2017 | 139.00 |
| 1 | Tuesday, February 7, 2017 | Thursday, July 20, 2017 | 163.00 |
| 2 | Friday, February 10, 2017 | | |
| 1 | Friday, February 10, 2017 | Thursday, March 16, 2017 | 34.00 |
| 1 | Friday, February 10, 2017 | Tuesday, April 18, 2017 | 67.00 |
| 1 | Monday, February 13, 2017 | Tuesday, August 1, 2017 | 169.00 |
| 1 | Monday, February 13, 2017 | Thursday, August 24, 2017 | 192.00 |
| 1 | Wednesday, February 15, 2017 | | |
| 1 | Wednesday, February 15, 2017 | Monday, May 8, 2017 | 82.00 |
| 1 | Thursday, February 16, 2017 | Monday, July 31, 2017 | 165.00 |
| 2 | Friday, February 17, 2017 | Tuesday, July 18, 2017 | 151.00 |
| 1 | Friday, February 17, 2017 | Thursday, July 27, 2017 | 160.00 |
| 1 | Friday, February 17, 2017 | Thursday, August 31, 2017 | 195.00 |
| 2 | Friday, February 17, 2017 | Thursday, October 19, 2017 | 244.00 |
| 1 | Tuesday, February 21, 2017 | Friday, May 19, 2017 | 87.00 |
| 1 | Tuesday, February 21, 2017 | Tuesday, August 8, 2017 | 168.00 |
| 2 | Tuesday, February 21, 2017 | Friday, November 10, 2017 | 262.00 |
| 1 | Thursday, February 23, 2017 | | |
| 1 | Friday, February 24, 2017 | Tuesday, April 18, 2017 | 53.00 |
| 1 | Thursday, March 2, 2017 | Tuesday, March 21, 2017 | 19.00 |
| 1 | Thursday, March 9, 2017 | | |
| 1 | Thursday, March 9, 2017 | Tuesday, June 27, 2017 | 110.00 |
| 1 | Thursday, March 9, 2017 | Wednesday, July 5, 2017 | 118.00 |
| 1 | Monday, March 13, 2017 | Thursday, November 16, 2017 | 248.00 |
| 1 | Tuesday, March 14, 2017 | Thursday, June 29, 2017 | 107.00 |
| 1 | Thursday, March 16, 2017 | Friday, December 1, 2017 | 260.00 |
| 1 | Friday, March 17, 2017 | Monday, July 17, 2017 | 122.00 |
| 1 | Monday, March 20, 2017 | Thursday, October 12, 2017 | 206.00 |
| 301 | | | 112.87 |

*Figure 16 Fourth tab of the dashboard with average lead times per category for clients coming through direct reporting*

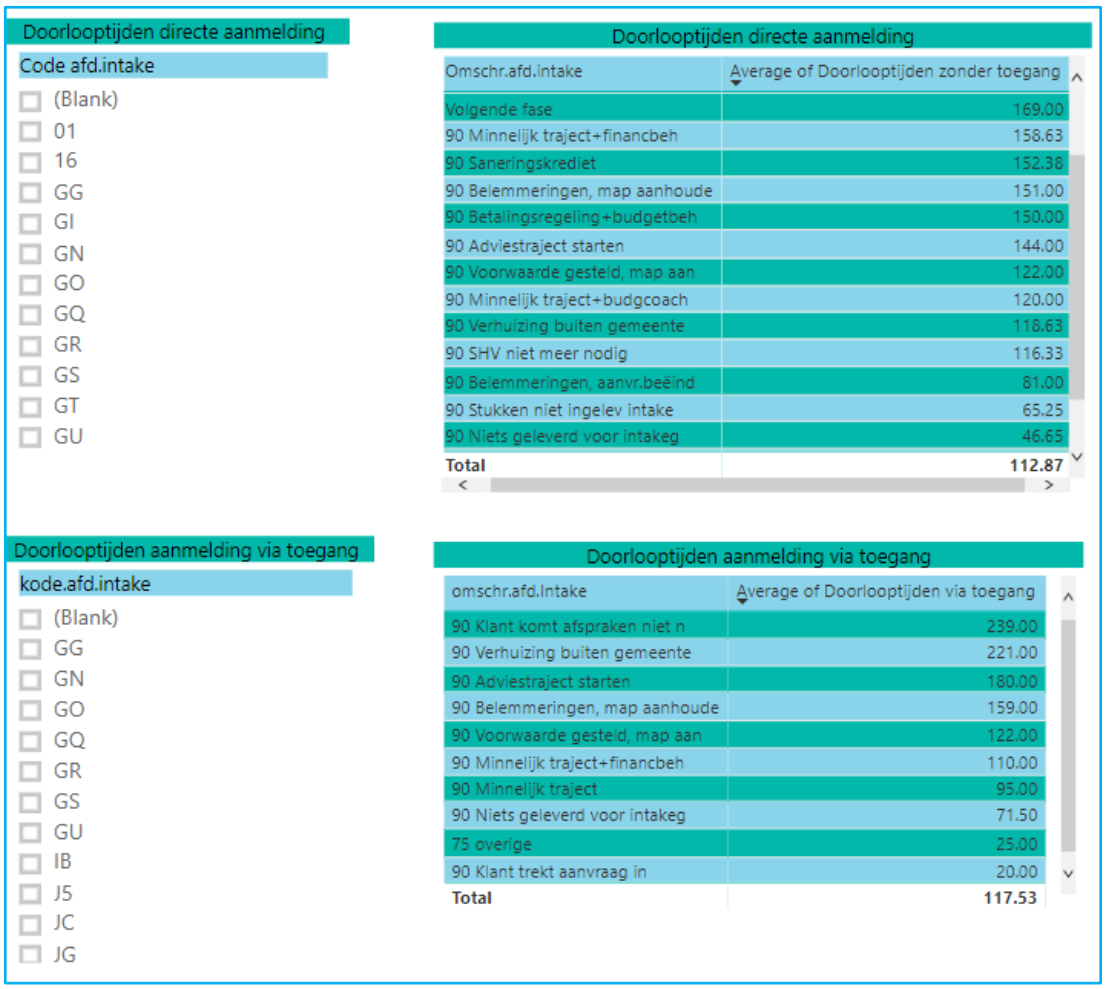



*Figure 17 Fifth tab of the dashboard where cases via the "toegang" and direct reported cases can be compared with each other by the decision codes*

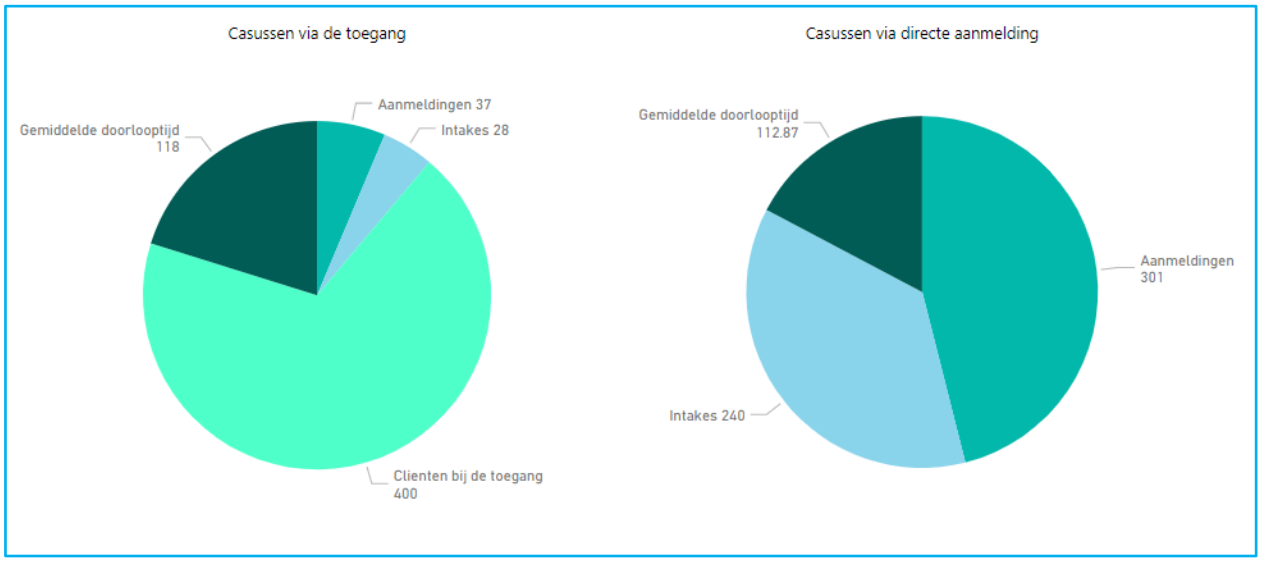



*Figure 18 Final tab of the dashboard with a summary*

Complete overview of second dashboard visualizations:

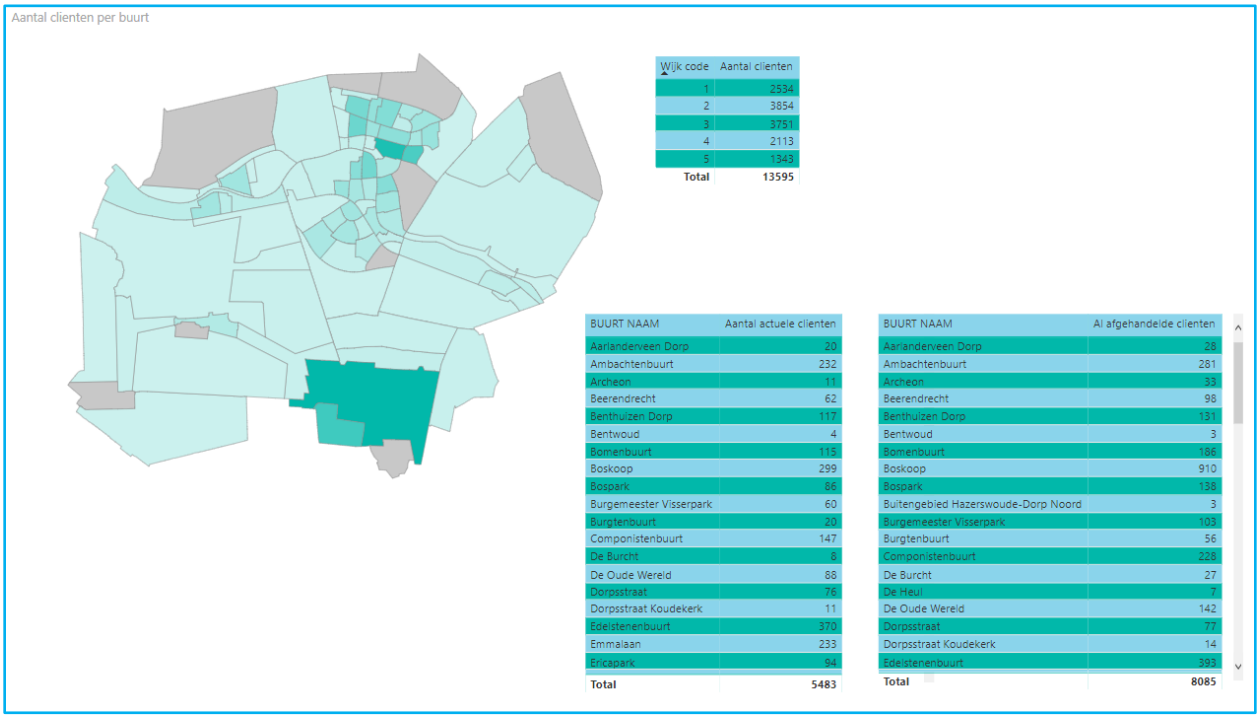



*Figure 19 First tab with information about active and closed cases categorized by district codes and district areas*

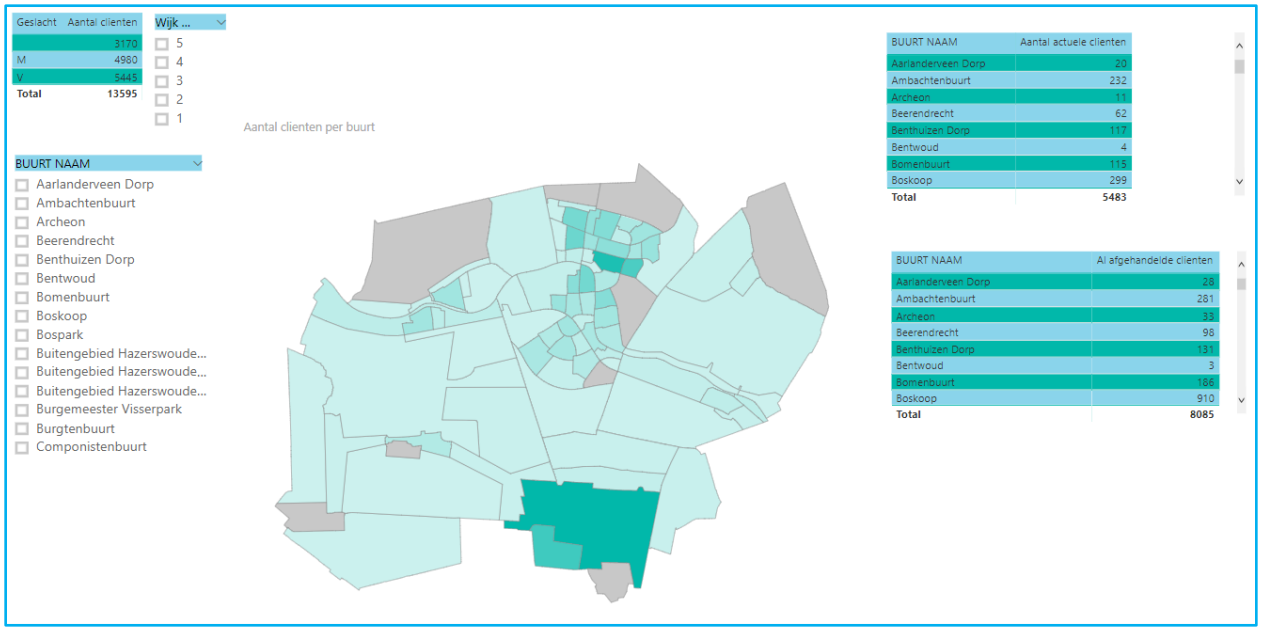



*Figure 20 Second tab with information about active and closed cases categorized by district codes, district areas and gender*

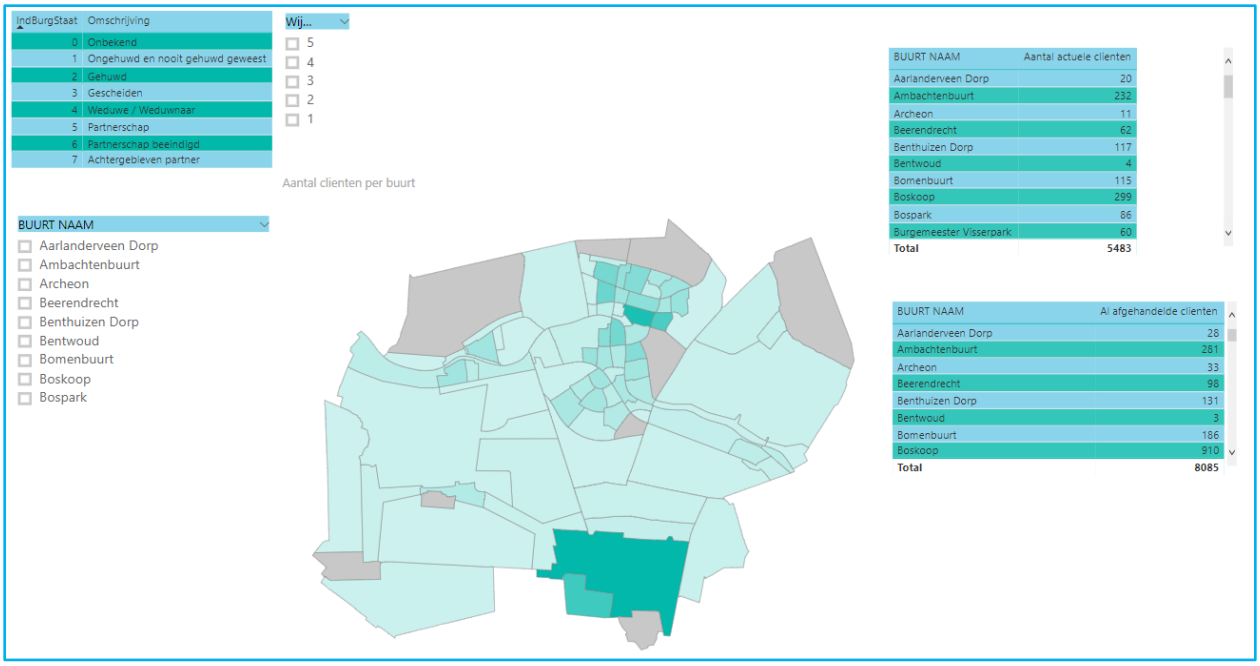



*Figure 21 Third tab with information about active and closed cases categorized by district codes, district areas and marital status*

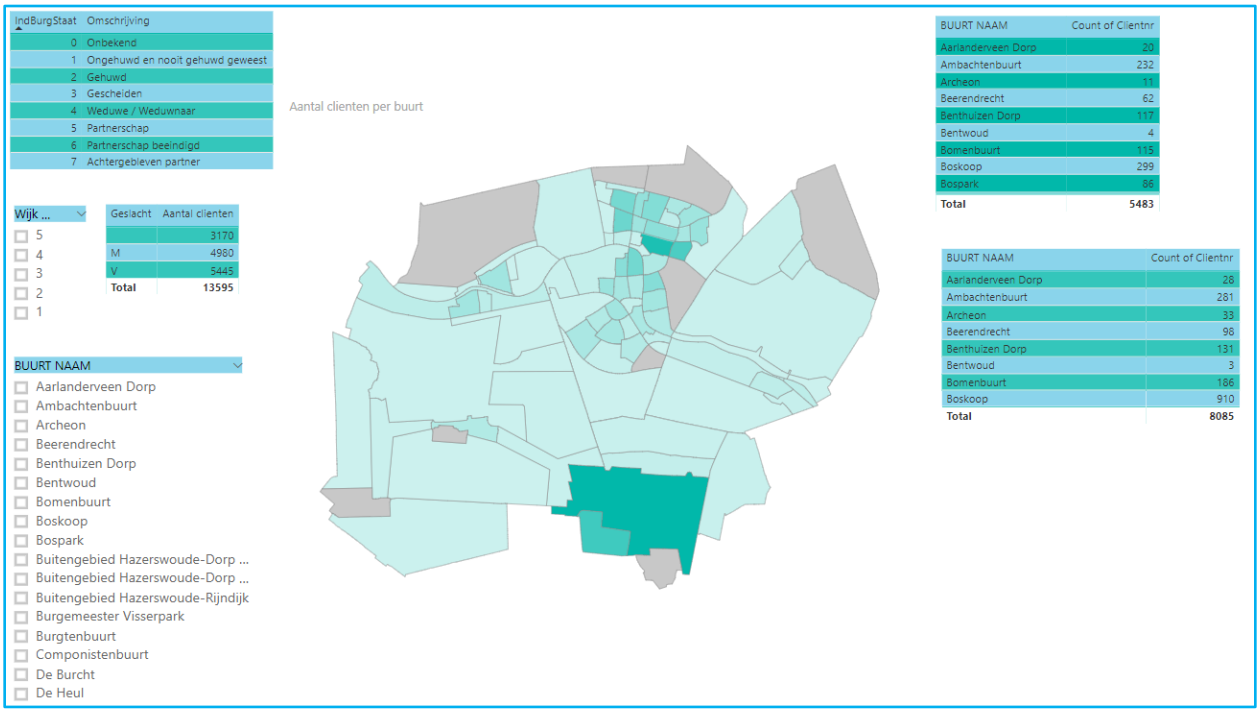



*Figure 22 Final tab with information about active and closed cases categorized by district codes, district areas, gender and marital status*

# Appendix B Content of Interviews and Results

**List of Interviewees**

Expert 1: Senior Data & Analytics Consultant at Motion10
Expert 2: Junior Data & Analytics Consultant at Motion10
Expert 3: Senior Technical Consultant Business Intelligence Motion10
Expert 4: Senior Database Administrator
Expert 5: Data platform project team leader / Financial Advisor for Engineering and City Management at the municipality
Expert 6: Senior Application Manager and Data Analyst
Expert 7: Functional and data manager at the information and automation department / ex senior IT consultant at Centric
Expert 8: Senior System manager
Expert 9: Privacy Officer / Legal Advisor
Expert 10: Advisor & Project manager for information management
Expert 11: Senior manager SHV department

*Table 10. Interview details*

| Type of interview | Group | Interviewees |
|---|---|---|
| Interviews with experts in the field of data visualization.<br>These experts were consultants with advanced experience in data visualization regarding various subjects in various environments (corporate as well as governmental). | 1 | Expert 1: Senior Data & Analytics Consultant at Motion10<br>Expert 2: Junior Data & Analytics Consultant at Motion10<br>Expert 3: Senior Technical Consultant Business Intelligence Motion10 |
| Interviews with representatives from the municipality of Alphen aan den Rijn<br>These representatives were members of the data platform project at the municipality. | 2 | Expert 4: Senior Database Administrator<br>Expert 5: Data platform project team leader / Financial Advisor for Engineering and City Management at the municipality<br>Expert 6: Senior Application Manager and Data Analyst<br>Expert 7: Functional and data manager at the information and automation department / ex senior IT consultant at Centric<br>Expert 8: Senior System manager |
| Interviews with the privacy officer of the municipality of Alphen aan den Rijn, who was also appointed as privacy officer by the municipality of Amsterdam during my research period. | 3 | Expert 9: Privacy Officer / Legal Advisor |

| | | |
|---|---|---|
| Interviews with employees of the department regarding my test cases | 4 | Expert 10: Advisor & Project manager for information management<br>Expert 11: Senior manager SHV department |

Questions asked during the interviews for determining the characteristics of the framework:

- What is your role in the organization?
- How often do you have to deal with sensitive information?
- Is sensitive data used for visualization in your organization?
- Why is visualization important in your organization?
- Why is sensitive data used for visualization?
- How is data stored in your organization?
- How is sensitive data visualized in your organization?
- What are the stages in the data visualization process?
- What is done to ensure privacy protection?
- Which privacy preserving data mining (PPDM) techniques are used during the data visualization process?
- Which rules and regulations need to be considered during the data visualization process?

Below you find a summary of the key findings during the interviews.

**How often do you have to deal with sensitive information?**
**Is sensitive data used for visualization in your organization?**
**Why is sensitive data used for visualization?**
All experts are confronted with sensitive information on daily basis.
The experts of group 1 are confronted with sensitive information from their clients that they have to visualize and present in the form of dashboards. The core business of this group is data visualization. For most of these visualizations private/sensitive information is used, because it helps the client to get more insight in his/her organization.
The experts of group 2 and 4 are confronted with sensitive information of the citizens of Alphen aan den Rijn. The municipality is moving towards digitization and during this process sensitive information is also used for the visualization. By using sensitive data more relevant information and insight can be achieved into the situation of the citizens of the municipality Alphen aan den Rijn. This information can be used for continuous monitoring and improvement of their organization.
And the expert from group 3 is also confronted with sensitive information of the citizens of the municipality Alphen aan den Rijn or the municipality of Amsterdam and other clients she offers legal advice. Throughout the process of digitization, it is one of her main tasks to see to it that the privacy aspects of the visualization process is taken care of.

**How is Data stored in your organization?**
Group 1: data in their organization is stored on windows servers. But they have experience with various other types of storage through their clients.
Group 2: data in their organization is stored on Oracle, SQL-Servers and hard copy dossiers.

Most data storage architectures can be divided into 3 main parts: the raw data, the modeled and modified data and the visualized data. These 3 main parts can be subdivided and may differ per organization.

**How is sensitive data visualized in your organization?**
**What are the stages in the data visualization process?**
For a start, it is necessary to know what you must visualize and why?
It is then necessary to know what information is available to achieve that, because for the visualization it is important that the data is digitally available.
The stages of the data visualization process are:

- Pulling the necessary data from the data sources to the data warehouse.
- Cleaning and modifying the data in a way that will be useful for further use.
- After that it is necessary to visualize the data in a way that fits the purpose of the visualization.
- Validate whether the result is accurate, privacy preserving and fits its purpose.

**What is done to ensure privacy protection?**
**Which privacy preserving data mining (PPDM) techniques are used during the data visualization process?**
To ensure privacy protection only necessary data is used. And to protect the sensitive and private data that is used during the visualization process various PPDM techniques are used. The mostly used techniques by Group 1 and Group 2 are anonymization, pseudonymization, aggregation and cloaking.

**Which rules and regulations need to be considered during the data visualization process?**
Every process that includes sensitive and private information needs to be GDPR compliant. The GDPR also encourages to use the anonymization and aggregation techniques.

Questions asked for validation of the framework:

1. Is the result of the visualization accurate? (Q.1)
2. Are the visualizations privacy preserving? (Q.2)

For testing the accuracy, the final numbers of the visualizations are matched with numbers the experts calculated themselves. These calculations were done with the applications they have been using till now. The created dashboards are meant to replace the old applications. The numbers were also compared to numbers presented by other dashboards regarding or partly related to the same subject. For testing the privacy preserving aspect of the visualizations the presented data was manipulated using different combinations of given information and investigated whether any combination represents sensitive information or information that can be used to eventually lead to a natural living person.

## Results

*Table 11. Results expert examination test case 1*

| Expert | Q.1 (Yes/No) | | | | | | | Q.2 (Yes/No) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Test case 1 | | | | | | | | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | y | y | y | y | y | y | y | y | y | y | y | y | y | y |
| 2 | y | y | y | y | y | y | y | y | y | y | y | y | y | y |
| 3 | y | y | y | y | y | y | y | y | y | y | y | y | y | y |
| 4 | y | y | y | y | y | y | y | y | y | y | y | y | y | y |
| 5 | y | y | y | y | y | y | y | y | y | y | y | y | y | y |
| 6 | y | y | y | y | y | y | y | y | y | y | y | y | y | y |
| 7 | y | y | y | y | y | y | y | y | y | y | y | y | y | y |
| 8 | y | y | y | y | y | y | y | y | y | y | y | y | y | y |
| 9 | - | - | - | - | - | - | - | y | y | y | y | y | y | y |
| 10 | y | y | y | y | y | y | y | y | y | y | y | y | y | y |
| 11 | y | y | y | y | y | y | y | y | y | y | y | y | y | y |

*Table 12. First results expert examination test case 2*

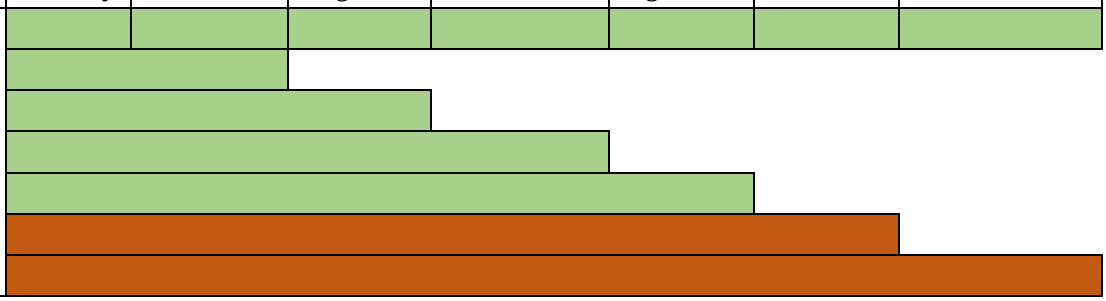
| Expert | Q.1 (Yes/No) | | | | Q.2 (Yes/No) | | | |
|---|---|---|---|---|---|---|---|---|
| | Test case 2 | | | | | | | |
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | y | y | y | y | n | n | n | n |
| 2 | y | y | y | y | n | n | n | n |
| 3 | y | y | y | y | n | n | n | n |
| 4 | y | y | y | y | n | n | n | n |
| 5 | y | y | y | y | n | n | n | n |
| 6 | y | y | y | y | n | n | n | n |
| 7 | y | y | y | y | n | n | n | n |
| 8 | y | y | y | y | n | n | n | n |
| 9 | - | - | - | - | n | n | n | n |
| 10 | y | y | y | y | n | n | n | n |
| 11 | y | y | y | y | n | n | n | n |

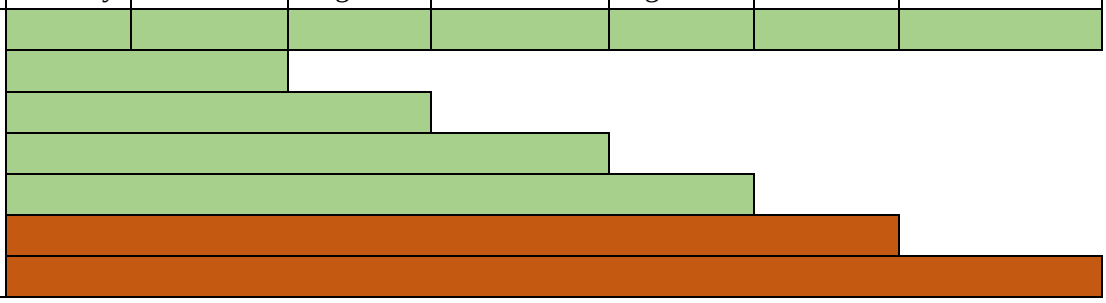
*Table 13. Second result expert examination test case 2*

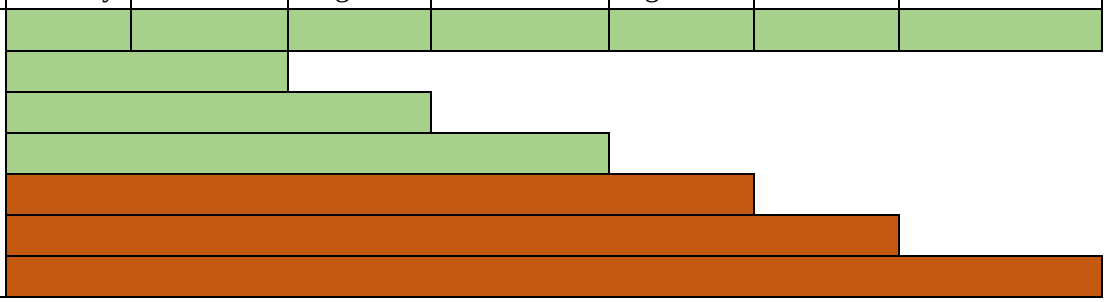
| Expert | Q.1 (Yes/No) | | | | Q.2 (Yes/No) | | | |
|---|---|---|---|---|---|---|---|---|
| | Test case 2 | | | | | | | |
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | y | y | y | y | y | y | y | y |
| 2 | y | y | y | y | y | y | y | y |
| 3 | y | y | y | y | y | y | y | y |
| 4 | y | y | y | y | y | y | y | y |
| 5 | y | y | y | y | y | y | y | y |
| 6 | y | y | y | y | y | y | y | y |
| 7 | y | y | y | y | y | y | y | y |
| 8 | y | y | y | y | y | y | y | y |
| 9 | - | - | - | - | y | y | y | y |
| 10 | y | y | y | y | y | y | y | y |
| 11 | y | y | y | y | y | y | y | y |

# Appendix C Data generation interface

# Appendix D Results of aggregation level

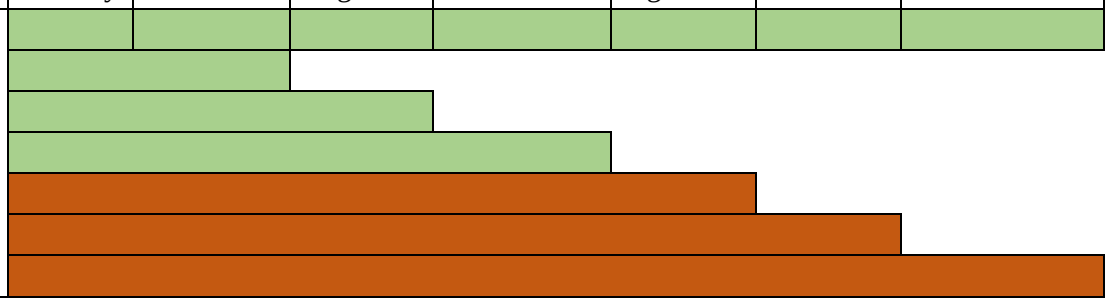
| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 90000 |  | | | | | | |

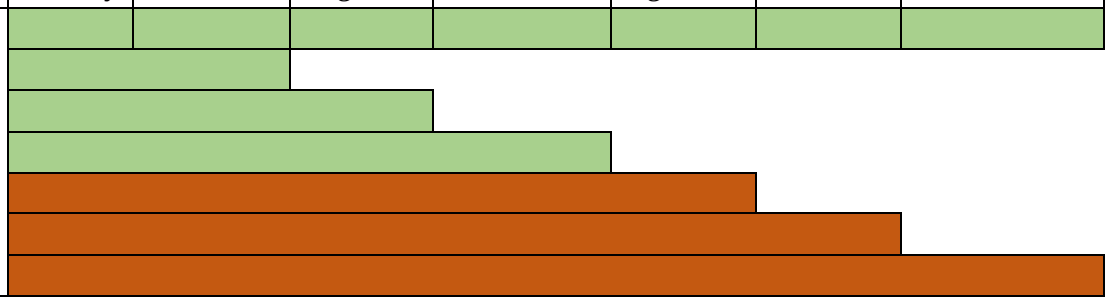
| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 75000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 60000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 45000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 30000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 20000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 15000 |  | | | | | | |

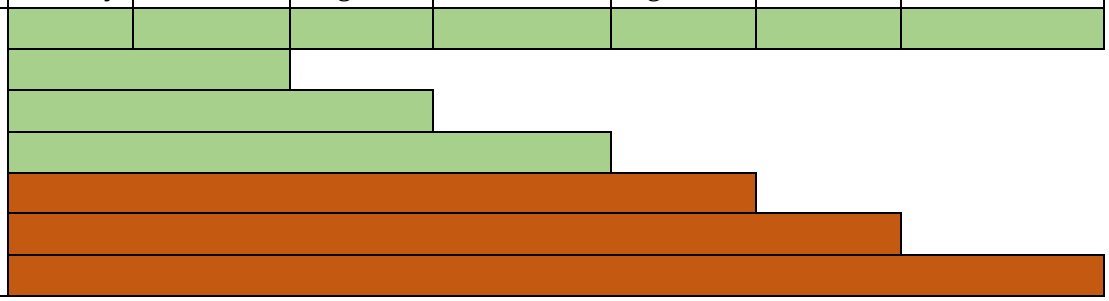
| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 10000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 9000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 8000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 7000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 6000 |  | | | | | | |

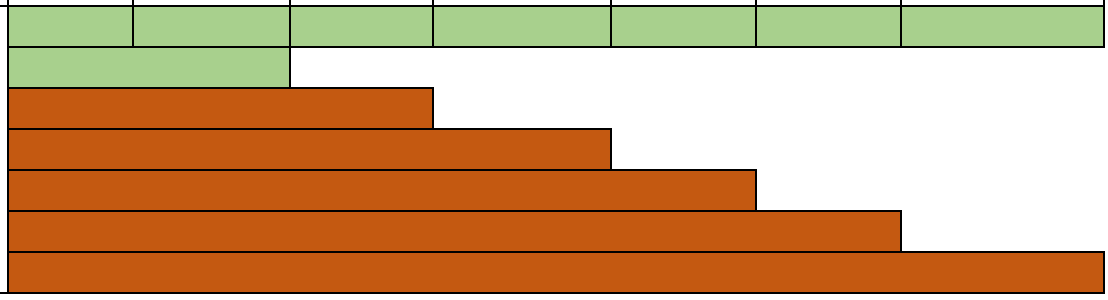
| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 5000 |  | | | | | | |

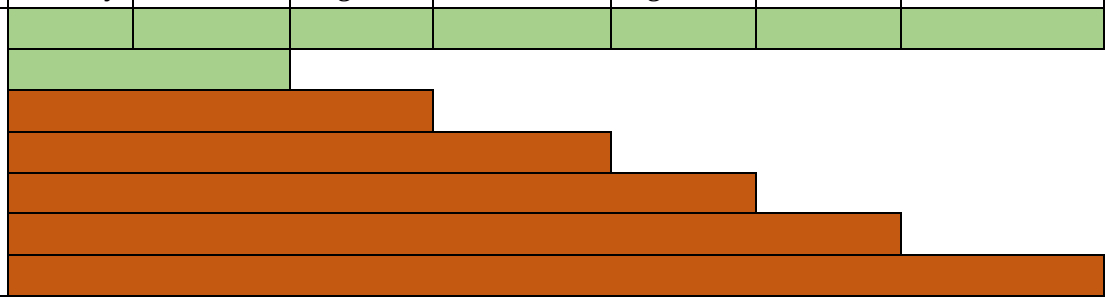
| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 4000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 3000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 2000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 1000 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 500 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 250 |  | | | | | | |

| Aggregation level | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Data set size | Country | Province | Region | Postal code | Age | Gender | Marital status |
| 100 |  | | | | | | |