Technische Universiteit Eindhoven

Visualization

2IV35

Visualization data of the Netherlands

Author:
Sander Kools
0848523
s.w.a.kools@student.tue.nl

Author:
Luuk Hulten
0720248
l.a.j.v.hulten@student.tue.nl

December 7, 2014

Contents

Information Visualization

In this report we will describe how we implemented a web application for visualizing a large data set for the course 2IV35. This data set contains information about the population living in the Netherlands.
There is a wide variate of data, for instance the living percentage for age ranges or car usage. The data is very large and it is hard to understand the data when viewed in the tabular view as it was provided, therefore we have come up with a better interface to make viewing and understanding the provided data easier.
In section 1, we will first give a description of the format of the data set.
In section 2, we will explain our design considerations for the interface.
In section 3, we will present our actual implementation, with screenshots and motivation.
In section 4, we will consider the visualization techniques that have not been used and why we choose not to use them but go for these instead.
Finally in section 5, we will show how our visualization can be usefull.

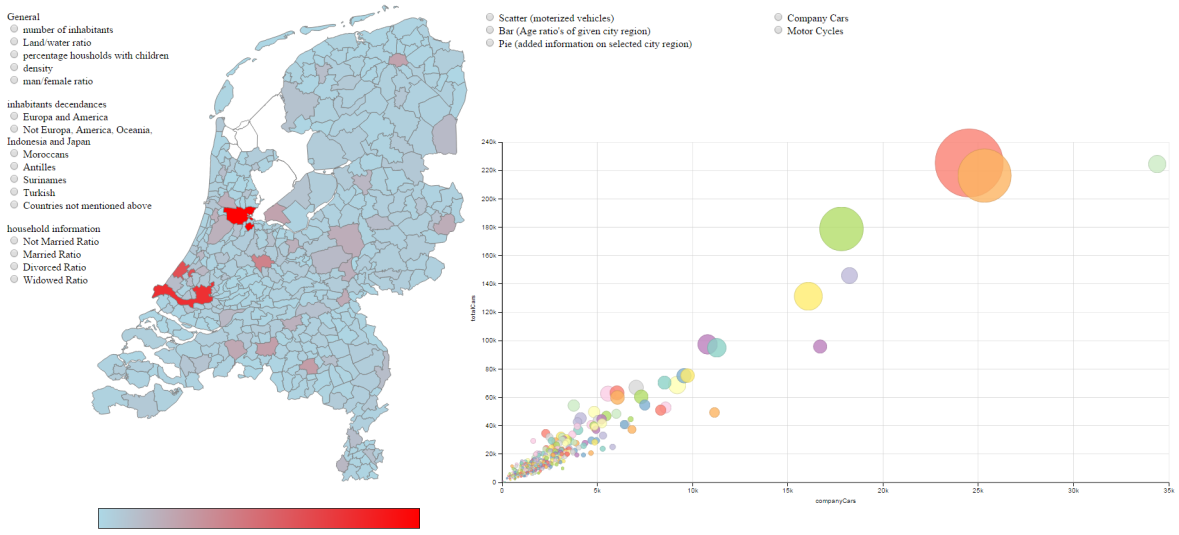


Figure 1: Overview of our application, showing two modes next to each other. Here we see the map of the Netherlands with population given in heatmap next to the scatter plot of total number of cars against company cars

1 Description of the Dataset

We have been given a dataset with information about the population living in the Netherlands. The data is provided in a .txt file with 417 rows and 60 columns every column is separated by a tab. Every row represents a city region inside of the Netherlands. This region is also describe inside a json file listing the coordinates of the region so that it can be drawn, all the regions combined represents the Netherlands.

1.1 High-level description

Our data set is called "cities-data.txt" and it describes various data about city regions in the Netherlands. The data is very detailed, for instance there is data given on the percentage of people between 0-14, 15-24, 25-44, 45-65 and 65 and older. but also a more detailed percentage is given of people between 0-4, 5-9 all the way to 95 and older.
The other data that is included is from the number of people and the density, surface area of land and water, descendance of the inhabitants, married and not married people, divorced people, widowed people, childeren in the household and number of cars, company cars or motorcycles. All the data is given for every city region. A detailed list of the data available can be seen in the subsection Description of the format found below.

1.2 Description of the format

The dataset originates from http://www.opencbs.nl (in Dutch), and it consists of shape information of Dutch municipalities in GEOJson format (cities-geometry.json), and a tab-separated file containing many statistics of these municipalities (cities-data.txt).
The following statistics are provided:

Column name	Description
GM_NAAM [string]	The official name of the municipality
WATER []	Has no relevance here
OAD [number]	Average number of addresses per square kilometer
STED [number]	Describes the urban character using the following classification: 1 very strongly urban 2500 addresses per km2 2 strongly urban 1500 2500 addresses per km2 3 moderately urban 1000 1500 addresses per km2 4 slightly urban 500 1000 addresses per km2 5 not urban ¡ 500 addresses per km2
AANT INW [number]	Number of inhabitants
AANT MAN [number]	Number of inhabitants
AANT VROUW [number]	Number of women
P 00 14 JR [percentage]	Percentage of inhibations aged 0 to 15 years
P 15 24 JR [percentage]	Percentage of inhibations aged 15 to 25 years
P 25 44 JR [percentage]	Percentage of inhibations aged 25 to 45 years
P 45 64 JR [percentage]	Percentage of inhibations aged 45 to 65 years

Column name	Description
P 65 EO JR [percentage]	Percentage of inhibations aged 65 years and older
P ONGEHUWD [percentage]	Percentage of unmarried people
P GEHUWD [percentage]	Percentage of married people
P GESCHEID [percentage]	Percentage of divorced people
P VERWEDUW [percentage]	Percentage of widows and widowers
BEV DICHTH [number]	Number of inhabitants per km2
AANTAL HH [number]	Number of households
P EENP HH [percentage]	Percentage of single households
P HH Z K [percentage]	Percentage of households without children
6P HH M K [percentage]	Percentage of households with children
GEM HH GR [number]	Average number of people in all households
P WEST AL [percentage]	Percentage of foreigners from Europe, North-America, Oceania, Indonesia, and Japan
P N W AL [percentage]	Percentage of foreigners not from Europe, North-America, Oceania, Indonesia, and Japan
P MAROKKO [percentage]	Percentage of foreigners from Morocco, Ifni, Spanish Sahara, and Western Sahara
P ANT ARU	Percentage of foreigners from the Dutch Antilles and Aruba
P SURINAM [percentage]	Percentage of foreigners from Surinam
P TURKIJE [percentage]	Percentage of foreigners from Turkey
P OVER NW [percentage]	Percentage of foreigners from other countries than mentioned in the above 4 attributes
AUTO TOT [number]	Number of cars
AUTO HH [number]	Number of cars per household
AUTO LAND [number]	Number of cars per km2
BEDR AUTO [number]	Number of company cars (minivans, trucks, etc
MOTOR 2W [number]	Number of motorcycles, including scooters
OPP TOT [number]	Total land and water area in hectares
OPP LAND [number]	Land area in hectares
OPP WATER [number]	Water area in hectares
P 00 04 JR [percentage]	Percentage of inhibations aged 0 to 5 years
P 05 09 JR [percentage]	Percentage of inhibations aged 5 to 10 years
P 10 14 JR [percentage]	Percentage of inhibations aged 10 to 15 years
P 15 19 JR [percentage]	Percentage of inhibations aged 15 to 20 years
P 20 24 JR [percentage]	Percentage of inhibations aged 20 to 25 years
P 25 29 JR [percentage]	Percentage of inhibations aged 25 to 30 years
P 30 34 JR [percentage]	Percentage of inhibations aged 30 to 35 years
P 35 39 JR [percentage]	Percentage of inhibations aged 35 to 40 years
P 40 44 JR [percentage]	Percentage of inhibations aged 40 to 45 years
P 45 49 JR [percentage]	Percentage of inhibations aged 45 to 50 years
P 50 54 JR [percentage]	Percentage of inhibations aged 50 to 55 years
P 55 59 JR [percentage]	Percentage of inhibations aged 55 to 60 years

Column name	Description
P 60 65 JR [percentage]	Percentage of inhibations aged 60 to 65 years
P 65 69 JR [percentage]	Percentage of inhibations aged 65 to 70 years
P 70 74 JR [percentage]	Percentage of inhibations aged 70 to 75 years
P 75 79 JR [percentage]	Percentage of inhibations aged 75 to 80 years
P 80 84 JR [percentage]	Percentage of inhibations aged 80 to 85 years
P 85 89 JR [percentage]	Percentage of inhibations aged 85 to 90 years
P 90 94 JR [percentage]	Percentage of inhibations aged 90 to 95 years
P 95 EO JR [percentage]	Percentage of inhibations aged 95 years and older

The data is delivered as a .txt file with columns that are separated with a tab. There is a lot of data to choose from, although a big part of the data is about the age range percentage. Also certain parts of the data can be determined using a combination of the other data available, for instance the "STED" can be determined using the "OAD" data or "OAD" itself could be determined using the surface area and the total inhabitants.

2 Design

In this section we explain the general design of the visualization.

2.1 General design

Since the dataset contains a large quantity of data about municipalities, a decision has to been made what to visualize. And also what visualization technique suits the selected data, so that it can be easily interpreted. More important is a valuable visualization, meaning that we can find valuable and useful information that meets the requirements.

2.2 Graphs

The dataset is mostly consists out of percentage so displaying this distribution in a pie chart seems obvious. In a pie chart you can easily see how a group of people is divided in subgroups. However a pie chart can not contain to much data since the sections become to small and hard to distinguish. In the case of the municipalities data married, divorced, widower and singles is an good example to represent in a pie chart. Since there are not to many categories and presented in percentage of the total population.
Also the distribution of the ages could be presented in a pie chart. But since these consists out of 20 categories, the pieces of the pie chart becomes to small and difficult to compare. Thats why for displaying this data a other representation needed to be found. Eventual a bar chart was chosen to display this data. A bar chart is capable of displaying many more categories. Another advantage of the bar chart is that you can easily compare data for multiple municipalities by placing multiple bars for each category with distinct colors for each municipality.
For finding the cars in contrast to the inhabitants for each municipality and finding outliers a scatter plot is a good choice. A scatter plot can contain lots of data, this enables us to plot all the car data of each municipalities at once in a single plot. Since the scatter plot needs to visualize three variables we extended the scatter plot. We made the scatter plot to represent the number of inhabitants by the size of the dots. We think that this enables us to find specific outliers very fast.

2.3 User Interface

There a many municipalities and there is a lot of data. Showing this at once for all municipalities makes the visualizations to complex and hard to inspect. Therefor we needed an interface that allows the user to scroll through the data and select some aspects of the data that needs to be visualized. To support these actions on the visualizations we needed an interface.
We started with a map of the Netherlands showing all municipalities. This map is used to scroll to the municipalities by clicking them. After clicking a municipalities the data selected is shown. Since we also want to make comparisons between municipalities, the ability to select two municipalities was added. When selecting two municipalities, the data of these two is added in one visualization. This approach makes it easy for the user to compare data between municipalities.
Since it is not useful to show all the data of a municipality at once, also an interface is needed to give the user to be ability to select some aspects about the data. In this case where there

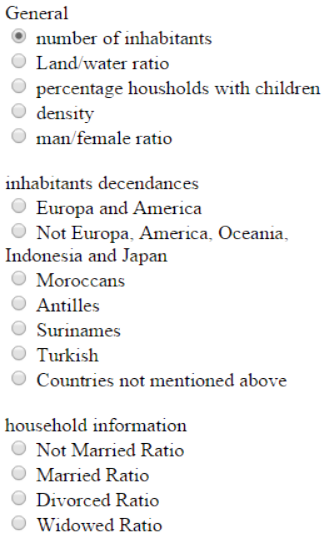
are not to many categories radio buttons is a good approach. Since they are small enough to place it direct in the visualization. This way the user gets a direct overview of the data and options available for the visualization without going through dropdown or list constructions. Selecting an aspect of the data would recolor the map to show an average overview of the data. This enables the user to find interesting municipalities and select them to get more detail of the data.
Details about the data is not only shown by creating different graphs but also annotating the graphs with the actual value. The advantage of annotating graphs is that misinterpreting the data is minimized. To keep the view clear and uncluttered the annotations are only shown when hovering over the elements.

2.4 Data Representation

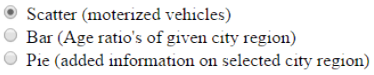
We tried to display as much useful and interesting data as we could from the cities-data.txt file, the data we don't use from the data as seen in the table is: **WATER, OAD, STED, AANTAL HH, P EENP HH, P HH Z K, AUTO HH and OPP TOT**.
Inside the data there were certain data aspects that were not given for certain city regions, for those entries the value became a big number like -99999997.0, also a row in the file had no function and could cause a misrepresentation of the data visualization. Because we didn't want to change the dataset given, we decided to identify the flaws and solve them in the coding of the visualization.

3 Implementation

We decided to to the implementation in HTML, CSS and JavaScript, the reasoning for this is because this way users will be able to visit the visualization easily using their web brower, without needing to install additional software.
For the visualization of data, we used the d3.js library. This library allows users to handle data in JavaScript easily. For the visualization we also used the scripts dimple.js and d3pie.js, these use the d3.js library and allow you to take full advantage of the power and flexibility of d3 to visualize data.
We had to use a range of visualizations to display useful data from the give dataset. For this we used the heatmap of the netherlands, a barchart, a scatter plot and a piechart.



(a) heatmap interface



(b) visualization interface

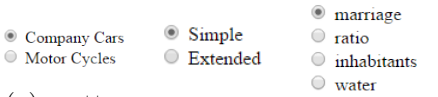
Figure 2: Radiobuttons

3.1 Technical Notes

We have tested the application in Firefox and Chrome. Other modern browsers like Safari, Opera and Internet Explorer should also support our application, but we have not tested that.
The visualization can be viewed by everyone, if you go to the site "www.isitsunday.com/2IV35.html" (use capital "IV") it will load the visualization we have made for the course 2IV35. How long the link will be supported can not be said at this time.

3.2 Main Interface

The interface that's implemented for this visualization is all realized with radiobutton's. You have the map of the Netherlands and with radiobutton's on the left side of the map you can make a decision of the data that is represented in the map (a). The map of the Netherlands will always be visible, so that means that the button's also will always be visible, so you are always able to change the data visualized in the map. There are also 3 radiobuttons on the top right of the map (b), these allow you to change the visualization that is visible on the left of the map, on default this is the scatter map, but you can change it to Bar chart visualization and pie chart visualization.
The radiobutton interface of the heatmap and of the visualization selection can be seen in the figure below and are visible at all times.



(a) scatter (b) barchart (c) piechart

Figure 3: visualization

The 3 visualizations that can be chosen between are scatter plots, bar charts and pie charts, for every chosen interface a new interface will become visible on the top left of that visualization for the 3 visualizations the interfaces can be seen in the figure below.

3.3 HeatMap of the Netherlands

For this assignment we were provided some basic code which would use the json file and draw the coordinates of the city region, which when combined represented the Netherlands. This has been modified in such a way that the city region changes colour when moused over and that every city region is clickable. The map will always be visible in the visualization, because the barchart and the piechart require input given from the map.
We have also included multiple datasets to be visualized in the heatmap. These have been split in 3 sections: General, inhabitants descendants and household information sectinos.
The General section includes: "number of inhabitants"(a), "land/water ratio"(b), "percentage households with children"(c), "density"(d) and "male/female ratio"(e).
The inhabitants descendants section includes: "Europe and America"(f), "Not Europe, America, Oceania, Indonesia and Japan"(g), "Maroccans"(h), "Antilles"(i), "Surinames"(j), "Turkish"(k) and "Countries not mentioned above"(l).
and the household information section includes: "not married ratio"(m), "married ratio"(n), "divorced ratio"(o) and "widowed ratio"(p). The letter after the data represents it's location in the figure displaying the heatmaps.
We chose this data to be the most interesting and we decided not to add more choices to the list because it would be to much and it would diminish the data that was represented. Data we considered to display but were not included were The Age range (a separate heatmap for every range) This has already been displayed in the barchart and the number of cars and motorcycles because it has been portrait in the scatter map.
All of the data that has been displayed in shown in the figure below.

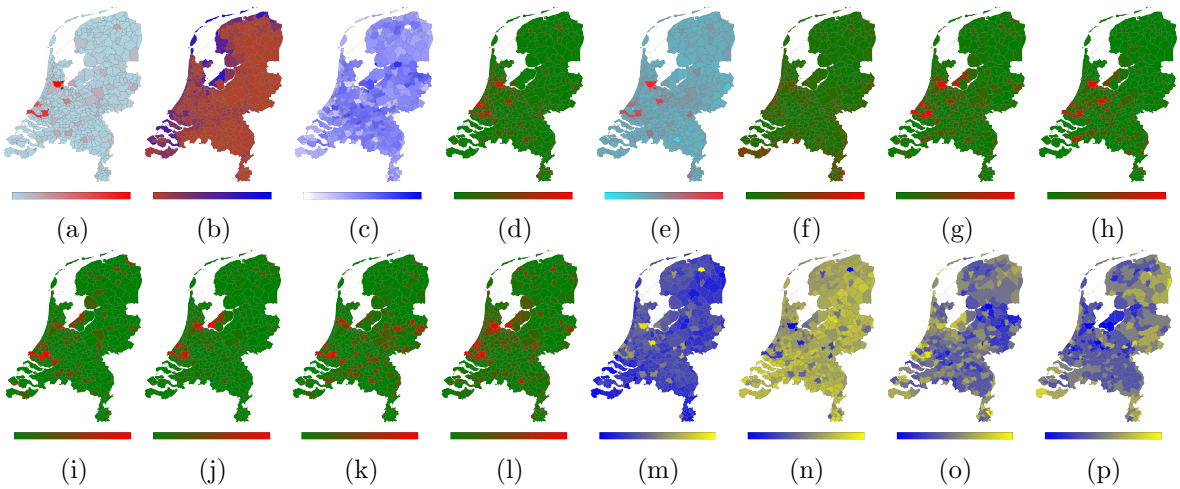
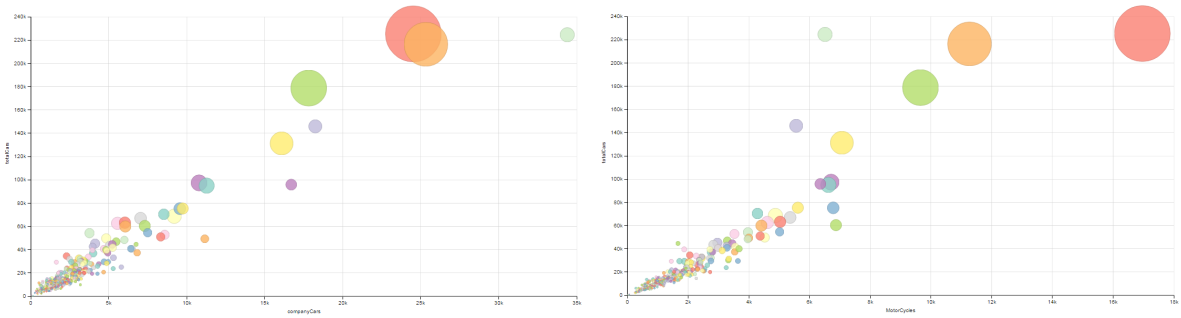


Figure 4: All Heatmap visualizations

3.4 Scatter Plot

The second visualization which on default is visible when running the site is a scatter plot, there are 2 options for the scatter plot as explained in the interface chapter, these options are company cars and motorcycles.
When selecting the company cars scatter plot, you will see the scatter plot of the total cars put against the number of company cars, every city region is a bubble and the size of the bubble represents the number of inhabitants. When selecting the Motorcycles the plot is similar, except instead of company cars, the number of motorcycles is displayed. You can mouseover a bubble, than a label will appear with the data of that bubble, so the name, inhabitants, total cars and total company cars/motorcycles. a line will be drawn across the x and y axis so you can better see at what point on the x and y axis the bubble is located.
The results of the 2 scatter plots can be seen in the figure below. with the company cars (a) and the motorcycles (b).

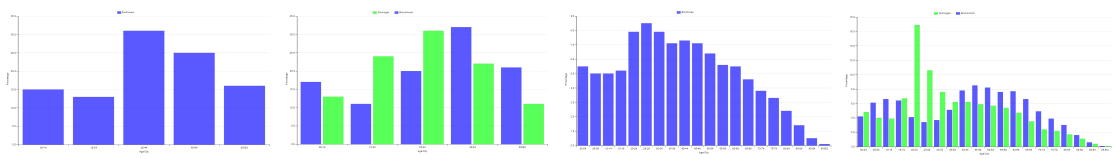


(a) company cars (b) motorcycles

Figure 5: Scatter plots

3.5 Bar Chart

The third visualization is a barchart where the age range percentage are displayed. You can select between simple and extended as described in the interface section. When the simple chart is selected you see the age range from 0-14 to 65 and older and with the extended chart selected you see the age range from 0 to 4, 5 to 9 all the way to 95 and older.



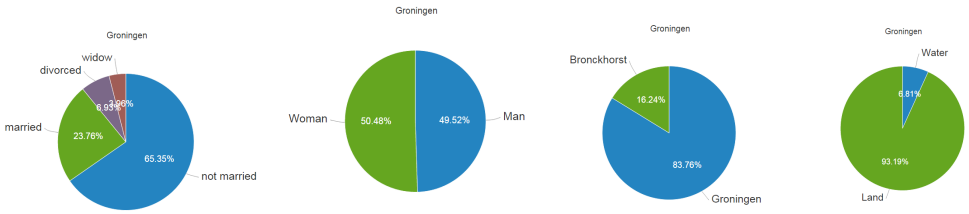
(a) small single (b) small extended (c) big single (d) big extended

Figure 6: Scatter plots

You give the bar chart input by clicking on the map, which is always visible on the left side of the screen, by clicking on a city region, the age range data of that region is loaded in the bar chart. When a second city region is selected both the data of that city region and the previous city region is loaded in the bar chart and they are given a different colour and a label to identify which colour is which city region. By double clicking a city region, you will see only the data of that city region.
So you can have 1 city region displayed with simple data (a) or extended data (c), or you can see 2 city regions next to each other with simple data (b) or with extended data (d) as seen in the figure.

3.6 Pie Chart

The fourth visualization is the pie chart where 4 different kind of data is visualized, the datasets that are displayed are kept small, because otherwise the pie chart would lose it's use. The first data that is visualized is the marriage data, so the percentage that is either married, not married, divorced or widowed. The second data that is visualized is the male/female ratio for that city region. The third data is the inhabitants ratio between the last 2 selected city regions. And the fourth data is the land/water ratio of the given cityregion.
Again you give input by clicking on the map that's always visible on the left side and the city region that is clicked will be the input for the data shown in the pie charts, also you can switch between the four data options by clicking the radio button's as described in the interface. In the figure you see an example of the married ratio pie chart (a), the male/female ratio pie chart (b), the inhabitants ratio pie chart (c) and the land/water ratio pie chart.



(a) Married (b) Male/Female ratio (c) Inhabitants (d) Water ratio

Figure 7: Pie Charts

3.7 Hovering and Selecting

To get more detail from a particular visualization, the user can hover over it with the mouse, this has an effect with every visualization.
With the heatmap hovering over a cityregion will colour it black, so you know for sure which will be selected when clicking, also if you hold the mouse still over a certain region, you will see the population in that cityregion and the name of the city region.
With the scatter plot you can mouse over the balls in the plot, doing this will drawn lines accross the x and y axis so you can easily determine both values. Also a label will appear

with all the information about that ball, so the name of the city region, inhabitants, total number of cars and the number of company cars/motorcycles.
With the bar chart you can mouseover the bars in the plot, the corresponding bar will have a line drawn to the y axis, so you can easily determine the value, also you can see the value in the label which will be shown when hovering over it, this gives the percentage of that bar and the given value of the x axis, with the name of the city region.
With the pie chart you can click on a pie section of the chart, this will cause the pie slice to pop out so you can beter determine the size, also the actual percentage is displayed in the slice.

4 Trends

We have used our application to analyze the data and found a couple of trends. By this we mean remarkable changes or values in the data.

The first trends we found was that there is a direct correlation between the number of inhabitants and cars within a municipality. However there is a obvious outlier and that is Almere, since it has lots of motor cycles in contrast to the number of inhabitants. As a reason we think that there are many lease company located there.

The second trend is that the most widowers are located near the borders of the Netherlands. We can not find a good reason for that. However if you compare the number of widowers to the age in a municipality you can easily find that the average age of these municipality is higher to the once without many widowers.

A third trend is that most foreigners live in the larger cities like Rotterdam, Amsterdam and Utrecht.

5 Conclusion

In this chapter we review the results of our visualizations.
The scatter plot is a good decision since it shows a lot of data in one view. But since most cities clutter in the left corner a part of the information is lost. We could improve this by making selecting different points possible and add zooming.
The bar chart displays the age categories very clearly. But also here we could improve by for example adding a average age line to it. An other improvement could be that instead of just clicking simple and extended view to unfold just a selected age bar in subcategories. This way you enable the users to directly search for the data.
Overall we think our visualizations suits the data good. We can easily find the answers of different questions. Nevertheless we think that the interface could be improved more. In some cases it is not directly clear what municipality is selected, since this is not reflected in the map. Also a clear flow of the application is missing. With this we mean that multiple views shows different information we could improve by showing the same information in different views. An approach like this gives the users the power to select the best view for the data instead of that the developer makes these discission for them. Naturally this is dependent on the target group of your application.

5.1 Testing

We tested our application with a select group of people and asked them to find the municipality with the most woman between the age of age of 20 to 25. They came up with Utrecht and Nijmegen as the best municipalities. So they successfully answered the question in a short amount of time. The test group was very interested in our visualization and liked to discover the data in the manner it was presented them.

Figure 8: Visualization testing group