Progress Report – CBS project

27 November 2020

university of groningen

DATA WISE
DATA SCIENCE IN SOCIETY

Group 11:

Augustin Procopi | s34842760 | a.procopi@student.rug.nl

Qihao Yi | s3414434 | q.yi.2@student.rug.nl

Raili Engler | s3140091 | r.engler@student.rug.nl

Rik Keppel | s3773787 | t.h.keppel@student.rug.nl

Tim Kramer | s2770660 | t.m.kramer@student.rug.nl

Wilbert van Renselaar | s3447294 | h.w.van.renselaar@student.rug.nl

Supervisor:

Arya Babai | a.babai@rug.nl

Content

I. Introduction

This report contains certain parts of progress made by group 11 on the CBS project regarding building a housing market dashboard. This project is part of the minor Data Wise at the University of Groningen. A project summary and several goals were documented in the project proposal. Here follows the project summary that was drawn up at the beginning of this project:

'The housing market in the Netherlands is generally experiencing growth, resulting in a housing shortage. In Groningen and the North of Drenthe the situation is different. Some municipalities are experiencing shrinkage resulting in a surplus housing supply, but others like the cities of Groningen and Assen are experiencing a shortage instead. The aim of this project is to gain insight in the reasons for these large differences in this small region'

Several goals were set, among others the theoretical learning part, the practical learning part and the deliverable for the CBS. In the past few weeks, the team worked really hard on those goals and significant progress has been made. In this progress report we outline and reflect on the progress made so far. Next to this, legal and ethical issues will be discussed. Also several professional issues and the benefits of our project will be analysed in the end. For certain parts of this report, a division is made according to the division of our project team, technical and literature.

As a final node, we are really glad and thankful with the help of our supervisor and project partner. From giving relevant suggestions to the feedback given, and from providing us with new ideas.

II. Progress of the Project

Selection of indicators

One of the project proposal deliverables is that a dashboard is delivered where several indicators are displayed that will be relevant for the housing market in the North of the Netherlands. To find the relevant indicators, a literature study was performed to find out which indicators are relevant to the housing market. While doing the literature research, it becomes clear that there is a large number of indicators influencing the housing market. Therefore, the choice was made to narrow down the scope of the research. The goal, according to the project proposal, is to get a better insight into the indicators that provides information about the shrinkage and growth of the different regions in the north of the Netherlands, and therefore it is logical to narrow down the scope of the literature research to the indicators that influence the shrinkage of regions in the north of the Netherlands. In addition, we searched for relevant literature that was partly provided by the project partner as well. This literature research was divided among the members of the literature group and an overview was made of the relevant indicators.

Out of this literature review, several indicators (approximately 20) were distilled and it is, due to time limitations, impossible to visualize all the founded indicators into our dashboard. Besides that, there was no data available that can describe these indicators. Therefore, a selection was made of three more objective indicators (i.e. that can be represented by registered data) and more subjective indicators (i.e. that can be represented by subjective data). The project partner advised us to make a distinction between registered data and subjective data. Registered data is more quantitative data where the opinion of people is not a factor of measurement, like numbers of houses or average income. Whereas subjective data is more focused on the collection of opinions of people, e.g. the satisfaction of people with the region of the liveability of their home. We choose to investigate both the registered and subjective data since we broaden our vision for both objective and subjective information regarding the topic of interest. The selected objective indicators were housing stock, housing price and migration within the region. The more subjective indicators are satisfaction with the liveability of the region, the perception of the vacancy rate of homes in the neighborhood and the willingness of people to move to another house.

The next step is to find the most relevant registered and subjective indicators. However, it seems to be quite difficult to rank the indicators on their relevance, since there is no literature available that compares and ranks the different indicators. But all the found indicators seem to have a significant impact on the shrinkage

of regions according to the reviewed literature. To make the selection, different factors are taken into account while narrowing down the indicators. Firstly, the frequency of the mentioned indicators is taken into account, i.e. how many reports considered the indicator as relevant (see Table 1). Secondly, the advice of the project partner was also considered. An example of the project's partner advice that changes our initial thought was the usage of the number of facilities as one of the indicators for shrinkage regions, i.e. a low number of facilities can indicate a shrinkage region. Rijnks (2020) finds out that the number of facilities is not a decisive factor for the shrinkage of regions in different regions . We also consulted other resources, like dashboards of the housing market abroad but also on a national level[1]. The last step made is the selection of the datasets that match the indicators. Statline is a website that provides all the registered open data of the CBS and this website is used to find data sets that describe the more objective indicators. The following three data sets are found on Statline: Housing stock, Average purchase price per region and the Population dynamics Statline does not provide the subjective data that is necessary that is needed for our research. The subjective data is provided by Woon Onderzoek. The Woon Onderzoek is a survey that is conducted by Dutch citizens of the and they are asked questions about their way of life, their satisfaction with their region, the type of house they live in, etc. (Dugteren et al., 2019). To select the data that matches the indicators, all the questions in the questionnaire were inspected to filter the relevant out of the irrelevant questions.

	Housing stock	Housing Price	Migration within the region	Satisfaction with the liveability in the region	Vacancy rate	Willingness to move
Nijskens & Lohaus (2019)	x	x	x	x		x
Dam et al. (2019)	x		x		x	x
Regio Groningen-Assen (2019)	x		x	x	x	

Table 1: The indicators that are specified as relevant in the literature

[1] Dashboards for inspiration: German dashboard, Housing market London, Housing price UK, Mortage rate USA, Home price USA

Indicators

The section above elaborates on the process of selecting the indicators that we will display in our dashboard. The next step in the process is a literature review for the selected indicators to get a better understanding of the indicators and the relationship between the indicators and the shrinkage of regions in general but also for the North of the Netherlands. First, the registered indicators will be discussed, followed by subjective indicators.

Registered indicators

1. Migration between regions

Dam et al. (2018) mentioned that there is a considerable amount of migration from the north of the Netherlands to the Randstad. The reason behind this phenomenon is maybe the increasing inequality between the wages, as Hamouda et al. (2012) states in his paper. The wages within the Randstad area tend to increase faster than other areas in the Netherlands. So it is reasonable that people move to this area searching for higher wages. As a result, wage inequality can be one of the deciding factors for people to migrate from the north of the Netherlands to Randstad.

Broersma and Oosterhaven, (2009) mentioned another explanation for the migration to the Randstad, namely that there is a positive relationship between the agglomeration effect and economic development in for example the Randstad. So the Randstad area tends to enjoy superior advantages in attracting migration by offering a variety of jobs as a result of the agglomeration of its multiple developed industries including examples like Phillips and KPMG. On other hand, Kempen et al. (2000) show that some people tend to move from city to rural regions because of for example unaffordable housing cost in the city. However, Heins et al. (2004) show that people will move from the city to rural areas but the rural areas nearby the city are preferred above the rural areas which are far from the city since the majority of jobs are within the cities. For the situation in northern Netherlands, the young people tend to come for education in Groningen and then move to the Randstad.(CBS, 2008).

To be more specific, the research from Broersma and Oosterhaven (2009) indicates that there is a positive relationship between the agglomeration effect and economic development. So the Randstad area tends to enjoy superior advantages in attracting the migrations by offering a variety of jobs as a result of the agglomeration of its multiple developed industries. On other hand, Kempen et al. (2000) showed that some people tend to move from city to rural regions because of for example unaffordable housing cost in the city. However, Heins et al. (2004) show that people will move from the city to rural areas but the rural areas

nearby the city are preferred above the rural areas which are far from the city since the majority of jobs are within the cities

2. Housing stock

There are two components of the housing stock, the first one is the countryside housing stock and the second one is the urban housing stock. Heins (2004) states in his report that the rural housing stock also has the requirements to protect the rural area from urbanization because of nature conservation. This can possibly lead to a shortage of housing stock in rural areas since the construction of new houses is limited. Heins also points out that this limitation is less important for the urban areas since there is less nature to conserve or protect.

The housing stock is also a factor that influences the satisfaction of people. Jansen (2013) shows that the increased housing supply will improve the satisfaction of the people in most cases. The satisfaction is improved since more people can have their own house .But the inequality among housing situations will decrease the satisfaction of the people that live in houses of lower quality. Therefore the quality of the available houses is also of importance besides the housing stock. For the situation in the Netherlands, Van Kempen et al. (2000) states that there is large inequality among the housing between different income groups in the Netherlands. The lower-income groups can be forced to move to rural areas because the housing prices are more affordable there. This can be an explanation for the migration from cities to rural regions.

When we come to the situation in the north of the Netherlands, there are different trends between the housing stock in rural areas compared to the cities. Sociaal Planbureau Groningen (2019) shows that the housing stock in the cities, e.g. Groningen, is increasing. Whereas the housing stock in the rural part of the shrinkage regions is decreasing. There is a small increase of the housing stock in the urban part of shrinkage regions. However, this is a significant smaller increase compared with cities in regions that are not exposed to shrinkage.

3. Housing price

The shrinkage of the region will directly affect the housing price. The research from Maennig(2008) states that there is a negative relationship between the shrink of the region and the housing price. The increase of the housing price on the other hand does not directly indicate economic development of the region because of the existence of housing bubbles. There are previous examples related to housing bubbles in developed

countries such as Japan (Cutts,1990). Cutts states that the increase of the housing price may not be the result of economic development but the result of government policy.

When we come to the supply and demand of the housing stock, it is stated that the change of the housing stock will have an immediate effect on the housing price (Wigren and Wilhelmsson, 2007). Wigren and Wilhekmsson also reveal that the change of the housing demand and supply has a lag-effect on the housing price, the change of the supply will have a comparatively faster influence on the housing price.

In addition to this, the housing price may also be correlated with the satisfaction of the owner of the house. The research from Ratcliffe(2015) states that there is a positive correlation between the housing price and the satisfaction of the owner. This relationship can even apply to the people renting the house when the housing price can represent the local development level. On the other hand, the satisfaction of the owner can also influence the housing price as Blair (2015) states in the report, there is a positive relationship between housing price and satisfaction of the resident.

Another important factor for the housing price is the density of the population. Florida (2016) found out that the housing price is higher in urban areas (i.e. higher population density) compared to rural areas (i.e. lower population density). This is especially the case for more developed countries like the Netherlands or the United States. This price gap between the rural and urban areas is growing even more the last few years.

Subjective indicators

1. Satisfaction with the region

Studies have shown that most people are in fact not satisfied with their current living situation (Crull, 1979). In the Britain at Home report (2020), Lloyds Bank Insurance reported on various housing market factors acquired through mass scale surveys. They found that circa 59 per cent of British residents are not happy with their current home. Of those 59 per cent, 26 per cent identified their dissatisfaction to stem from the fact that they are not living in their desired area. Similarly Terry's Fabrics survey (Shotton-Pugh, 2019) that took place throughout Britain also reported that more than a third of British homeowners fell out of love with their home as soon as they moved in, with 25 per cent feeling that they should have taken more time to decide whether the property was right for them.

Most studies investigate residents' general satisfaction with their home, including the actual house, the neighborhood and region. Residential satisfaction is the subjective evaluation of various house, neighborhood and regional characteristics. Neighborhood satisfaction has however been shown to be an

essential independent predictor of housing satisfaction (Abidin et al., 2019). The indicator that is going to be focused on in this research project is residents' satisfaction solely with the living environment in the region. This is because housing style and quality can vary greatly and patterns could potentially get lost when looking at entire regions rather than streets or houses.

Neighborhood satisfaction is an important predictor of residential satisfaction as a whole (Abidin et al., 2019). One of the main predictors of neighborhood satisfaction is the distance travelled to medical centers, schools, shopping, work etc (Abidin et al., 2019). Quality of streets and roads, homogeneity regarding social class, race, and ethnic group, crime rates and the physical environment also factor into the reported neighborhood satisfaction (Abidin et al., 2019).

2. Vacancy rate of homes in the neighborhood

In the Netherlands more than 400 thousand houses are unoccupied and 190 thousand remain unoccupied for more than 18 months (Statistics Netherlands, 2014). An increasing vacancy rate indicates that the region would indicate shrinking. Nonetheless, due to the decline in population, an increasing vacancy of housing will be inevitable (PBL Netherlands Environmental Assessment Agency, 2010) Residents of the area may notice this trend when the streets and neighbourhood gets emptier. Stores, medical centers and schools may also reduce in number. Location is a primary factor people take into consideration when deciding on which property to buy (Abidin et al., 2019). The easy mobility and closeness of facilities provided by cities has attracted large parts of the global population for the past decades (Abidin et al., 2019). Therefore, one may expect that if people notice the vacancy rate of homes in their neighbourhood rising, they plan differently for the future as they foresee facilities and mobility to also reduce. If regular citizens have noticed an increased vacancy rate, private investors and builders are probably also less likely to purchase properties in the area, which could halt the housing market all together. Nonetheless, it is also possible that in some regions the vacancy rate is high but the construction of new buildings is also high, as residents prefer to live in newly built homes. Those residents that do decide to move away from the increasingly vacant region would likely move to a growing region, such as cities and suburbs.

3. Willingness to move

Mobility as defined by Speare in his 1974 study on residential satisfaction can be measured by the wish to move to a new home in the first year and subsequent mobility one year later. He found the wish to move in the first year to be positively related to subsequent mobility. Naturally, satisfaction with the initial home was negatively related to wish to move and subsequent mobility. Residential satisfaction, wish to move, duration of residence, and home ownership had significant independent effects on subsequent mobility.

However, together these factors only explained 24 per cent of the variance in residential mobility, while residential satisfaction and home ownership explained 25 per cent of the variance in residents' wish to move.

This clearly raises the question of what factors into the variables in order to explain the other circa 75 percent of the mobility and wish to move variables. Why do all factors together only account for 25 per cent of the variance in residential mobility? One may consider that the wish to move is not the same as the propensity to move, which would indicate that the residents have desires, plans, inclinations or expectations about future mobility (Morris et al., 1976). However even if some people would like to move but think it is out of reach or are not seriously considering it. Why does residential satisfaction have a relatively small impact on the wish to move? One may contemplate whether people often move spontaneously, whose intention or desire to move is not caught by the 'wish to move' variable. They may be happy with their house, however stumble across one that they feel would suit them better, and spontaneously move.

Additional factors that may factor into the intention or desire to move are financial constraints, family and personal circumstances, housing stock and various other indicators also considered in this project. One should also consider that in most studies on this topic, residential mobility includes any move, even if it is just down the block. The shrinkage regions in the North of the Netherlands considered in our project are larger regions and are concerned with mobility over longer distances than within local neighborhoods. Nonetheless, the general trend that arose from the literature on this topic indicates that the willingness to move is a significant predictor of subsequent changes in address. We expect to see this positive correlation in the North of the Netherlands as well.

Tech Group

One of the deliverables that was promised in the project proposal is a dashboard that visualizes several housing market indicators. To make a dashboard is not a trivial task: it requires automated data collection, data transformation and data visualization. It requires programming skills to make these processes automated. Fortunately several members of our project group indicated they were able to program. These three members set out to implement these automated processes and make the dashboard a reality. In the project proposal it was uncertain which data, which indicators or which visualization techniques would be used. This resulted in the schedule being rather vague and general: the tech group was supposed to 'set up infrastructure', 'program' and its leader was supposed to 'coordinate technical team' up to this point in the project. The task to set up infrastructure was the clearest at the time of writing the project proposal. This

task was planned to be completed in the first week (October 18th to 24th). The programming and coordinating was planned to start the week after.

The following learning activities were completed by the technical team:

1. Exercise on using Git, R package structure and automated testing
2. Exercise on using GitHub

The following infrastructure activities were completed by the technical team:

1. Set up a repository on GitHub
2. Set up a R package structure in the repository
3. Set up unit testing framework using testthat
4. Set up Continuous Integration using travis-ci
5. Set up code coverage checks using codecov.io

The following dashboard related activities were completed by the technical team:

1. Extracted housing stock data for the North of the Netherlands
2. Cleaned housing stock data for the North of the Netherlands
3. Implemented basic R line graph to plot housing stock data over the period 2015-2019
4. Fixed a namespace conflict between the packages ‘jsonlite’ and ‘shiny’
5. Added aesthetic layout elements to the dashboard to make it look nicer and more professional
6. Implemented a more visually appealing and interactive line graph
7. Extracted housing stock data for the North of the Netherlands per municipality
8. Cleaned housing stock data for the North of the Netherlandsper municipality
9. Implemented a dropdown menu to select which municipality to show the housing stock line graph over the period 2015-2019 for
10. Implemented a sidebar to select other (to be implemented) indicators
11. Implemented a data caching function
12. Fixed an R version related bug in automated tests
13. Fixed automated tests failing due to a package throwing depreciation warnings
14. Extracted housing price data for the North of the Netherlands per municipality
15. Implemented housing price graph for the North of the Netherlands per municipality

The following dashboard related activities have been started by the technical team:

1. Implement a choropleth map of the North of the Netherlands per municipality
2. Impute housing price data for municipalities existing in 2019 as if they had existed over the period 2015-2019

The aim of the technical team is still the same as mentioned in the project proposal: To realize a dashboard on the housing market in the north of the Netherlands.Our methods have stayed the same as well: we are using the R language and the R package ‘shiny’ as web framework to make this dashboard a reality.

III. Reflection on Progress

Project Proposal

The previous section describes the progress that is made so far, especially it pointed out the process of finding the indicators for our dashboard. One of the first things we performed was limiting the scope of the project. The topic of shrinkage regions was chosen since there are multiple municipalities in the north of the Netherlands that deal with this issue. To illustrate the shrinkage of some regions, Meier (2014) shows that all municipalities in Drenthe are subject to shrinkage, except for regio Assen; the other two northern provinces also have a few municipalities that experience this shrinkage. It was a challenge for our group to find the right balance between a research goal that is too broad or too limited within the given time frame of the project. The first draft that we delivered for our research proposal was based on the housing market in the north of the Netherlands in general. With the feedback from the project partner and the supervisor we came to the conclusion that the chosen scope for the project was too general and as a result, it is hard to know where to start. We could have prevented this situation by discussing our ideas with the project partner and the supervisor at an earlier stage. Fortunately, the first draft was handed in a week before the final deadline and therefore we had enough time to change the scope of the project for the project proposal. The communication of our ideas is improved after this event and we will definitely continue this practice in the upcoming stages of the project. An example of this is the discussion of the relevant indicators with our project partner in an early stage and therefore we could implement his advice before the actual work was done yet.

Furthermore, there was another issue with the set goals in the first draft of the project proposal. One of the goals was to deliver a dashboard that policymakers can use in their decision-making and thereby improve the housing market in the northern Netherlands, especially in shrinkage regions. This goal is really ambitious since it is unknown if we will build a dashboard with valuable and actionable information for policymakers. The supervisor also emphasized that it is important that we do not arouse too high expectations about our project towards the project partner, e.g. propose realistic goals instead of too ambitious goals. This was an eyeopener for our group and therefore we decided to make the main goal of our project more realistic and durable. We changed the main goal in ‘help our project partner, CBS, by realizing a housing market dashboard’. Also in this situation, we can indirectly contribute to the provision of relevant information to policy makers.

According to the project proposal, it was also the aim to work on the deficiencies of knowledge about programming. The team members of the tech group spend a significant amount of time to get a better understanding of the programming language and the limitations and opportunities of the program. However, the team members of the literature group do not dive into the programming language and therefore there is a lack of understanding about the code that is written for the dashboard. The literature group did not have the time to get a better understanding of the code. This phenomenon leads to a lack of understanding by the literature group regarding the work of the tech group and therefore it is not always clear what the possibilities and limitations are for the dashboard, e.g. the way data can be visualized. The lack of programming knowledge is partly because of time constraints due to the electives of the minor. Since these electives are finished, there is more time for the project and therefore the members of the literature group will follow some programming classes on data camp to decrease the deficiencies and get a better understanding of the work of the tech group.

Indicators

The listing of relevant indicators was the step after the limitation of the scope for the project. Different papers and reports are viewed to detect the indicators. However, it is also good to be aware of certain biases in this stage. Hao (2019) elaborates on the bias of the ‘unknown unknowns’ (i.e. you do not know what you do not know). When a literature review is done to find indicators, you only find the indicators that are present in the reviewed literature but maybe other relevant indicators are missing in this literature. The literature review is based on different sources to prevent substantial unknown unknowns, nevertheless, it is hard to remove the unknown unknowns with certainty. So it is a challenge for our project to elaborate more on the validity of our indicators and it is also important that we have a certain openness for new relevant indicators that are not present in our study yet.

Another aspect of the used literature is the extent to which the context of the report is applicable for our project. Zook et al. (2017) also address the importance of the use of context as well. The consideration of the context is relevant since not all our literature was written in the context of our project. Only the article by Van Dam (2009) is written in almost the same context, namely the shrinkage regions in the North compared with the city of Groningen. However, this article is also 11 years old and the context of 11 years ago is not necessarily the context of today. The other sources are not specifically written for the shrinkage regions in the North of the Netherlands but provide information about shrinkage regions in general. This was an important point of consideration for our project and we discussed and explained to each other why we think that a certain indicator fits in our context or not. This discussion is not reported yet, so it is very

important to transparently report this thinking process in the final report by arguing why the chosen indicators also fit for research in the North of the Netherlands.

The use of dutch and foreign housing monitors is a good example where the context is an important point for consideration. These housing monitors are used as inspiration for the selection of possible indicators and the way how we can visualize them. However, the housing monitors are mainly built to give a better impression of the housing market in general and are not especially devoted to providing information about shrinkage regions in general or for the north of the Netherlands. Therefore we cannot use the information of the dashboard directly but we have to filter out the information that fits the context of our project. This is partly done with background knowledge we obtained by studying literature about the topic and the filtering is also done with the discussion of information with the other team members. This discussion and considerations are not reported yet and this is maybe a point for improvement since the reporting is likely more accurate if it is performed in the short term.

The context of the used data is also relevant to take into account while using it. One of the data sets is the Woon Onderzoek To ensure that the context of this data is suitable for usage, different aspects of the data creation are investigated. The collection and representativeness of the respondents are for example investigated by taking a look at the clarification documents that are available for the Woon Onderzoek. In these documents, it was stated that oversampling was used to make the data representative. Hauner et al. (2014) investigated the oversampling method on potential biases and errors and he did not detect a bias or substantial error in the oversampling method. Besides that, Hauner et al. argue that oversampling can be a useful tool to use to make the data representative. For this research, it was concluded that the Woon Onderzoek data is contextually relevant since it provides data about the regions we are investigating with ensuring the representativeness of this data by the use of oversampling.

Technical part

There were two main challenges to making the dashboard a reality which could be alleviated by proper infrastructure. First, there was the challenge of working together. Source code (the product of programming) is very sensitive to changes in structure. Where a certain line of code is placed matters, because the program may be interpreted in a different way. Even a misplaced character may cause an error. When working together, it is therefore incredibly important to have a very clear structure (what goes where?) and to merge the progress made by all collaborators often. The clear structure prevents collaborators from working on different things in the same file or section of a file, while the frequent merging ensures that any ambiguity regarding what parts should go where are found and resolved as quickly as possible.

The clear structure is provided by our choice to make an R package out of our dashboard. R packages are collections of code other people can download, install and use in their own projects. Examples are the package ‘dplyr’, which makes transforming datasets easier, and ‘shiny’, which we are using as a web framework to make our dashboard. Packages have a certain file structure, where all R code in the package is encapsulated in functions, which are all in separate files in the package’s ‘R/’ folder. This encapsulation in functions and separate files makes it so when collaborator A is working on function X and collaborator B is working on function Y, they (should) never have any conflicts when merging their code. A’s code is in a different file and in a different function than B’s.

As an added benefit, using a R package structure allows us to easily run automated unit tests. These tests check that the functions behave as expected in at least a couple of circumstances. As an example, imagine a function ‘add’ that takes two arguments and adds them together. A test for add would be that it returns 4 when it is provided the arguments 2 and 2. As you can see, this testing is not exhaustive, but it is a check that the code does what it is supposed to do. It therefore helps improve the code quality and works as documentation, showing exactly how the original writer of the function expected the function to behave. We incorporated unit testing in our project using the package ‘testthat’.

To enable and encourage frequent merging, the version control system Git was used. This system allows for tracking of when, what and why files in the version-controlled repository were changed. In very basic terms, a collaborator downloads the repository from a central server, makes changes, documents those changes, uploads the changes to the central server and then requests the changes they made to be merged into the main repository. As long as collaborators work on small issues at a time, this enables all collaborators to be up-to-date and have the most recent code changes available. For this project we are using the platform GitHub to provide us with the central server. This platform provides an accessible web-based user interface to track issues, requests to merge (called pull requests on GitHub) and discuss those changes with the rest of the team.

The second challenge that could be alleviated by infrastructure is the challenge of ensuring a base level of code quality. Code quality is understood as the code doing what it is supposed to do and it being clear to people reading the code or its documentation what it is supposed to do. As none of our project group members are highly experienced programmers, having checks in place to check the code does what it is supposed to do is imperative. It enables the technical team to quickly find bugs or mistakes in the code, which makes it less likely that weeks are spent finding the cause of a single strange bug that is somewhere

in an undocumented thousand-line source code file. Secondly, it makes the code in this project more reusable and therefore more valuable to the project partner.

To ensure a base level of code quality, unit testing (explained above), continuous integration, code coverage checks and pull request reviews are used. Continuous integration is an automated system that checks whenever a change is made or proposed to the central repository if the repository still conforms to the R package structure and if all unit tests pass. A similar automated process is run after the continuous integration system finds no issues: the code coverage check. This checks if the unit tests actually test all of the source code that has been added to this (proposed) change. If no new tests have been added, but new code has been added, the code coverage check will report that less code is covered by tests. This information provided by the continuous integration system and the code coverage check is integrated with the pull request information provided to the pull request reviewer. This enables the reviewer, which is a different member of the technical team than the one that proposed the change, to see if the code has been properly documented, does what it is supposed to do and it is understood by another person what the code is supposed to do.

During the first week it became apparent that the infrastructure choices had an additional cost. Because not all members of the technical team were familiar with the version control system (namely Git), automated unit testing (facilitated by the R package testthat) and the dashboard framework (provided by the R package shiny), some extra time was needed to learn how these things work. Some exercises were created by the technical team leader to facilitate and structure this process. This did however delay when the team was able to start programming. There was a shift in team composition as well, as described in the professional issues chapter of this report. Lastly, the technical team leader did not need to spend all of his time coordinating the group, which meant he could join the programming effort.

SWOT analysis on Collaboration Process

We carried out a SWOT analysis of the collaboration process within our group. We identified the aspects that we are strong at and already bring to the table and the situations in which we can use them in order to excel at a task. Additionally, we discussed our group’s weaknesses that we will be confronted with in situations of threat. This analysis is a valuable exercise for this group project as it gives us an opportunity to reflect on what parts we are strong at and what we need to work on. Situations in which weaknesses can become a problem for our progress can be prepared for and prevented.

Strengths	**Weaknesses**
- Good mix of people and skills (Belbin) - Motivated team members that are willing to invest time in the project - Clear task division - Good communication with project partner and supervisor	- COVID-19 measures - Busy schedules and work beside the project
Opportunities	**Threats**
- Improve collaboration skills, especially in an online environment - Learn to efficiently and professionally communicate with external parties	- Meetings need to be planned in advance - Do not limit the scope of the project in a sufficient way

Table 2: SWOT on Collaboration Process

The Belbin team role exercise identified the strength within our group that there is a mix of people from different cultural and professional backgrounds and that we bring a diverse set of skills to the table. All team members are also very motivated and are willing to invest time and effort into the project. Based on our interests, skills and previous experience, the task division was clear. The group was divided into two teams, the technical team and the literature team, who each know what their contribution should be. Regularly scheduled meetings with the supervisor and project partner have ensured good communication. We have a clear idea of what they would each like to see from our group and we frequently update them on our progress. This is also an opportunity for us to improve upon our skills in communicating with external parties and ensuring all people related to this project are satisfied with the way the project moves forward. This project also allows us to experiment with completing group work in an online environment as it is new to all of us. Nonetheless, the COVID-19 measures remain a weakness of our collaboration process, as we are not able to all come together physically. Clear task division and online communication are all the more important in this time. Various group members also have jobs, courses and other engagements on the side of this project and their busy schedules could be considered a weakness. We try to plan around that by scheduling our meetings with appropriate advancements. Nonetheless, a threat the project group is faced with is that when an important pressing issue comes up, we cannot (all) meet. The housing market and the

task of programming a housing monitor is new to the whole team. It is of importance that we do not lose sight of the time, resources and skills we have. We should ensure that the scope of the paper only covers the material that we are able to deliver in a high-quality and timely manner. Overall, the collaboration process has made it clear that the threats and weaknesses can be overcome by planning ahead.

IV. Timetable

The timetable of the remaining weeks can be found on the shared drive with the supervisor and project partner. This schedule is a copy of the original schedule, which will be updated weekly. The schedule shows that we have got enough time remaining to finish the goals set for the project. Even some buffer time is incorporated in the final weeks.

V. Project issues

Legal issues

On the 24th of May 2018, the GDPR was introduced by the European Union (EU). According to Voigt & Von dem Bussche (2017), the GDPR has a major impact on the field of data research. The privacy rights of human data objects are ensured with a variety of regulations. Since it is necessary for our project to make use of data it is important to know what is legal and illegal to do with the data that is available for our project.

The GDPR provides the data subjects with a number of rights that are relevant to take into account while handling the data; these rights are the following according to the European Parliament and Council of European Union (2016):

- Right to information (Article 13 and 14): Data objects have the right to be informed about the procession and usage of their data
- Right of access (Article 15): The data subject has the right to know if his data is processed and where and who is processing it. Besides that, they have the right to access their own personal data.
- Right to rectification (Article 16): The data subject has the right to rectification if the personal data of the data subject is incorrect.
- Right to the restriction of processing (Article 18): The data subject has the right to limit the use of data under special circumstances.
- Right to object (Article 21): The data subject has the right to object if the personal data is shared with third parties.

It is essential to investigate the legal basis on which the registered and the WoonOnderzoek data is collected by CBS and therefore the data policy of CBS will be discussed regarding the stated articles above. Firstly, the European Parliament and Council of European Union (2016) shows in article 44 an exemption of article 15,16 and 18 for data processing that is performed by institutions that will use the data for scientific or statistical purposes. Since, CBS is an institutional organization that collects the data for statistical purposes it is exempt from articles 15, 16, and 18. The CBS (2020a) also elaborates on their own website about the legal rights of their data subjects. The rights that are described in articles 13 and 14 are ensured by informing the data subjects of what is done with the collected data. Article 21 is not applicable for CBS, since CBS ensures that personable trackable data is not made publicly available. CBS (2020b) shows data microdata

can be made available to third parties. However, the created output of third parties is controlled by third parties to ensure that the published data is not personally traceable.

Another legal aspect that is related to the data we use is the license that is used by CBS to license their data. CBS (2017) uses a Creative Commons Attribute license. This means that it is required to refer to the used dataset to make clear that the data is originally from CBS, but not that the impression is given that CBS agrees with your work or your use of the work. Therefore we have to explicitly state that the data is provided by CBS but also make clear that the procession and presentation of the data is our responsibility.

Ethical issues

Besides the legal issues, there are also ethical issues that are considered during the project. The Dutch Code of Conduct for Research Integrity by KNAW et al. (2018) is used to identify and discuss possible ethical issues that are related to our project. This code is binding for all the Dutch universities and therefore also for this project since the project is performed by students from the University of Groningen. KNAW et al. (2018) make use of five principles which are the basis for scientific integrity: honesty, scrupulousness, transparency, independence, and responsibility.

There are some factors that can possibly conflict with one of these principles. For example that the project partner CBS wishes one kind of outcome for our project. This can influence our independent research and honesty because there is potentially a threat that our research results are adapted to satisfy the project partner or that unwelcome findings are neglected. However, this is not a big issue for our project, since the project partner is legally binded by the CBS regulations. CBS (2020c) states that the CBS will never draw conclusions or make prognoses when they are working with other parties. This means that the project partner is legally not able to influence the outcome of the results. And if the outcomes are influenced, we can submit complaints to Statistics Netherlands, which will then be dealt with (CBS, 2020d). Fortunately, we do not expect this situation since our project partner provides us with advice but he never tries to influence the outcomes of the research.

Another aspect that needs to be taken into account is the cleaning of the data. This needs to be done in an honest, scrupulous and transparent way. To use the data for our research, it is necessary to clean the data to make it for example machine-readable but also to eliminate some disturbing factors in the data, e.g. the merge of some municipalities. There is a risk that relevant data is cleaned and that therefore the dataset is

not ready for use. To prevent this from happening, it is necessary to argue and explain why some data can be cleaned or why not. Besides that, the reason why data is cleaned, why it was possible to clean the data and how the data is cleaned needs to be reported to maintain transparency about the usage of the data.

Professional issues

For our project, several professional issues were identified in the first few weeks of the project. In this chapter, those issues will be identified, elaborately explained and ways to mitigate those issues will be discussed. Later on, several risks to the project partner will be examined.

To start with, in the beginning of the project we found out that our group was quite diverse when it comes to skills available but also talking on cultural diversity. The professional issue of every team member working partly on every part of the project did not seem like a good idea to us. Therefore, after 2 weeks already, the team was divided into two subgroups. Namely the technical and the literature group, in order to let this work efficiently, the difference of tasks between those groups were discussed. Two weeks later, it appeared that one team member did not feel at his place in the subgroup he was placed in. To solve this problem, a meeting was scheduled with two people from each subgroup and several tasks were contrived to let the team member feel helpful. Since his available skills were overlapping between the tasks of both teams, the missing key was found. For the team members of both subgroups it was sometimes unclear what the other teams were working on. Luckily, by letting this team member being the bridge between those subgroups, it could be really seen as a win-win situation for all team members.

The role division and the assessment of different roles available was also supported by the course ‘Dynamics of Multi-disciplinary Teamwork’. During this course, several papers were read about the importance of balanced groups and the effect of team charters. We made a Belbin role division, because one of the most important claims in Belbin’s work is that balanced groups have superior performance than unbalanced groups (Belbin, 1981). It can be said according to the different team roles, that our team is really balanced in terms of different roles available and therefore Meslec and Curseu (2015) indicates this positively predicts group performance in preliminary phases of the group project. Also the significant relevance of the team charter we had to make is supported by Aaron, McDowell and Herdman (2014), who found support for the assertion that the introduction of team charters does in fact manifest improved process outcomes, including communication, effort, cohesion, mutual support and member satisfaction.

Another issue was the lacking programming skills of several team members. Since the minor we choose is data wise, all team members are willing to learn and work on their programming skills. However, in our

team it was clear from the beginning that some team members were way more experienced than others. To solve this problem and to let everyone become more ‘data wise’, one of our team members came with the idea of following online trainings. The platform we used for this was DataCamp, first because it was a platform where most of the courses offered were free of charge. Second, they had two specific courses relevant for us, introduction to R and intermediate R, which was a combined workload of approximately 10 hours.

As a group of students with all different disciplines, we had a broad range of knowledge available. From technical studies to sociology and from business economics to mathematics. However, none of us has a background in real estate studies or something related. Since the housing market is complex and consisting of many different indicators and variables, our team had a theoretical backlog in the beginning. To overcome this issue, the literature group worked hard on finding relevant papers, reading rapports about the housing market, searching for existing dashboards etc. We were glad about the help of our project partner, who was more experienced in this field and already collected certain reports that were specified on the North of the Netherlands. By reading, studying, and summarizing for the technical group, our knowledge regarding the housing market expanded over the weeks.

The following issue did not only hit our project team, but all project teams over the world. COVID-19. Because our team was careful in handling the covid-19 measures, physical meetings were scarce. This influenced the dynamics within the team. Collaboration online has its advantages, but for projects with this size, physical meetings are required so now and then. Unfortunately, the university had a policy where it was not allowed to rent classrooms for project teams, even though most of them were empty all the time. Luckily one of our team members was doing a board-year somewhere, therefore several meetings were held in that building. The last consequence of the virus on our team is that one of our team members is going back to his home country by the end of November. This decision is fully understandable on the one hand but could influence the team spirit of our group on the other.

Several risks to the project partner can be drawn from the professional issues stated above. Although we applied multiple ideas of overcoming certain issues, some risks of our project are inevitable. First, since our knowledge regarding the housing market keeps limited, the risk of using misleading information or variables in our dashboard remains. The same reason can be used for the second risk, since we are students from different backgrounds and none of us has ever built a dashboard before. The risk remains that the final product we deliver does not reach the standards we had in mind. An example could be that the monitor does not offer enough information and new insights. Also the outcome of the indicators analysed can be different from the existing literature because other methods and techniques were used. These risks are minimized by

discussing our steps and choices with the project partner and supervisor and besides that we try to have a broad literature overview where we do not focus on one particular part of the literature available. In the following section the benefits of our project will be discussed for society, the economy and the scientific benefits. Finally, the potential risks and benefits will be balanced to prove that our project is of big value.

VI. Balancing Risks and Benefits

We have identified several potential benefits that our final products can deliver. There are 3 main aspects of these benefits. The first one is the **societal** aspect, and we believe that there are two main agents who will find the dashboard we are creating useful.

- The *policy makers* have to constantly manage and improve the housing market. Their decisions cannot be based only on theoretical knowledge, they need to analyze vast amounts of data over several periods in time that can show trends in a specific area of this market on specific regions, in order to propose efficient housing policies (Aldridge, 2018). Even though policy makers responsible for the northern part of the Netherlands have access to this data, in its current state it is scattered on more platforms, it is not represented in an intuitive way and filtering through this data can be tedious. Our dashboard can be considered as a subset of all registered and subjective data regarding the housing market in the north of the Netherlands that is available, and it has only the main factors required by the policy makers to make informed decisions, as identified by our research. Moreover, the visualizations in the dashboard are meant to make this data more understandable and intuitive.
- Providing a clear overview of the housing market can be useful for *citizens* too. With our registered data, the housing price indicator allows citizens to make a more informed decision regarding buying a house or investing in real estate.With our subjective data, the general satisfaction of citizens living in certain neighbourhoods can be used by potential buyers to decide what regions might suit their preferences more. Having at least this basic information about the housing market might be enough to make the purchasing decision without a real estate broker.

The second one is the **economical** aspect, which obviously has tangencies with the societal one, but we will only mention how our project can be used by private sector agents in order to have a clear distinction on the main aspects of the benefits. Housing market characteristics of a region are usually taken into account by some companies that want to activate either in that region or in a neighbouring one.

- For example, businesses that require low-skilled workers, such as factories or businesses from the horeca industry, might be interested in the dynamics of the population of the city they want to activate in. If they have to choose between two regions, one of which shows a continuous shrinkage in the past 5 years and a decreasing housing stock, and the second one shows the complete opposite,

then logically they would choose the second region as there might be a higher number of potential workers.

- Another example would be the real estate agencies. Decisions on which regions to invest in can be based both on the subjective and the registered data provided in our dashboard. Take a region with the following characteristics: average selling prices are relatively high, they keep increasing and a lot of people migrate in. These characteristics might be enough for these companies to invest in such a region.

The last aspect we want to present is the **scientific** one. Our literature team does an extensive research on reports, papers and national surveys related to the housing market in the Netherlands in order to find the most important characteristics that directors in the region need to make use of for their activity. We believe that having this general overview of the housing market might find indicators that are of utmost importance but have not been researched enough as they have an indirect impact. Moreover, the analysis of our selected indicators and their graphical representations can be seen as an input to the already existing research.

Several issues were identified in the legal, ethical and professional issues part, the most important ones were wrong analysis of the public data, presenting our own work as work from the CBS and making wrong assumptions when cleaning the data which can lead to the absence of relevant information. In order to mitigate those risks and issues, several arguments were already given. It keeps really important that the methodology of our progress should be reported precisely. Next to this, if we clean the data or make certain assumptions, enough research should be provided to justify those assumptions. Lastly, when doubting about certain issues or being at extra risk in some situations, collaboration with the project partner and supervisor is really important. They are experienced in the field of doing research and know when something is legitimated.

We are strongly aware of the risks we face in this project. As already mentioned, we do everything we can to mitigate those risks and solve the issues. Therefore the benefits outweigh the risks and our project can be a great contribution to either society, the economy and for scientific research.

VII. References

Aaron, J. R., McDowell, W. C., & Herdman, A. O. (2014). The effects of a team charter on student team behaviors. Journal of Education for Business, 89(2), 90-97.

Abidin, N. Z., Abdullah, M. I., Basrah, N., & Alias, M. N. (2019). Residential Satisfaction: Literature Review and A Conceptual Framework. *IOP Conference Series: Earth and Environmental Science*, *385*, 012040. https://doi.org/10.1088/1755-1315/385/1/012040

Aldridge, S. (2018). The housing market: challenges and policy responses. National Institute Economic Review, 245(1), 8. https://doi.org/10.1177/002795011824500110

Broersma, L., & Oosterhaven, J. (2009). Regional labor productivity in the Netherlands: Evidence of agglomeration and congestion effects. Journal of Regional Science, 49(3), 483-511.

Belbin, M. (1981). Managenmt Teanu. London: Heinemann.

Blair, J. P., & Larsen, J. E. (2010). Satisfaction with neighbors and neighborhood housing prices. Journal of Place Management and Development.

CBS. (2008). De stad Groningen als roltrap van Noord-Nederland. Statistics Netherlands. Available at: https://www.cbs.nl./nl-nl/achtergrond/2008/28/de-stad-groningen-als-roltrap-van-noordnederland(Accessed: 26 November 2020)

CBS. (2017). Disclaimer Open Data CBS. Statistics Netherlands

CBS. (2020a). Privacy. Statistics Netherlands. Available at: https://www.cbs.nl/en-gb/about-us/organisation/privacy (Accessed: 14 November 2020)

CBS. (2020b). Microdata: Zelf onderzoek doen. Statistics Netherlands. Available at: https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen (Accessed: 24 November 2020)

CBS. (2020c). Nieuwe regelgeving geeft richting aan de rol van het CBS. Statistics Netherlands. Available at: https://www.cbs.nl/nl-nl/corporate/2020/27/nieuwe-regelgeving-geeft-richting-aan de-rol-van-het-cbs (Accessed: 25 November 2020)

CBS. (2020d). Klachtenregeling in het kader van de toepassing van de Beleidsregel taakuitoefening CBS, Regeling werkzaamheden derden CBS en het instellen van een klachtencommissie. Available at: https://www.cbs.nl/nl-nl/corporate/2020/27/nieuwe-regelgeving-geeft-richting-aan-de-rol-van-het-cbs (Accessed: 25 November 2020)

Cutts, R. J. (2014, August 1). Power from the Ground Up: Japan's Land Bubble. Harvard Business Review. Available at: https://hbr.org/1990/05/power-from-the-ground-up-japans-land-bubble. (Accessed: 26 November 2020)

Dam van, K., Wissen van, L., Roo de, G., Dijk van, J., Strijker, D., & Veenstra, J. (2009). Regionale bevolkingskrimp en de Stad Groningen. Quick Scan. Faculteit Ruimtelijke Wetenschappen, Groningen.

Crull, S. R. (1979). *Residential satisfaction, propensity to move, and residential mobility: A causal model* (Doctor dissertation). Iowa State University. https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=8197&context=rtd#:~:text=Residential%20satisfaction%20and%20home%20ownership,moves%20within%20a%20local%20area.

European Parliament and Council of European Union (2016) Regulation (EU) 2016/679. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN (Accessed: 13 November 2020)

Florida, R. (2016). The Incredible Rise of Urban Real Estate. Bloomberg. Available at: https://www.bloomberg.com/news/articles/2016-02-25/the-growth-of-urban-vs-suburban-housin prices-zillow-report (Accessed: 26 November 2020)

Hamouda, O. F., de Groot, H. L., & Groot, S. P. (2012). Wage inequality in the Netherlands: Evidence, trends, and explanations. Journal of Income Distribution.

Hao, K. (2019). This is how AI bias really happens—and why it's so hard to fix. MIT Technology Review.

Hauner, K. K., Zinbarg, R. E., & Revelle, W. (2014). A latent variable model approach to estimating systematic bias in the oversampling method. Behavior research methods, 46(3), 786-797.

Heins, S. (2004). Rural living in city and countryside: Demand and supply in the Netherlands. Journal of Housing and the Built Environment, 19(4), 391-408.

Jansen, S. J. (2013). Why is housing always satisfactory? A study into the impact of preference and experience on housing appreciation. Social Indicators Research, 113(3), 785-805.

Kempen, van, R., Schutjens, V. A., & Weesep, van , J. (2000). Housing and social fragmentation in the Netherlands. Housing Studies, 15(4), 505-531.

Lloyds Banking Group. (2020). *Britain at Home Report*. https://www.lloydsbankinggroup.com/globalassets/documents/media/pressreleases/lloydsbank/2013/3009_home_report.pdf

Maennig, W., & Dust, L. (2008). Shrinking and growing metropolitan areas asymmetric real estate price reactions?: The case of German single-family houses. Regional Science and Urban Economics, 38(1), 63-69.

Meier, S. (2014). Demografische veranderingen in Noord-Nederland.

Meslec, N., & Curşeu, P. L. (2015). Are balanced groups better? Belbin roles in collaborative learning groups. Learning and Individual Differences, 39, 81-88.

Morris, E. W., Crull, S. R., & Winter, M. (1976). Housing Norms, Housing Satisfaction and the Propensity to Move. *Journal of Marriage and the Family*, *38*(2), 309. https://doi.org/10.2307/350390

PBL Netherlands Environmental Assessment Agency. (2010). *From combating to managing: Demographic decline in the Netherlands*. PBL Publishers. https://www.pbl.nl/sites/default/files/downloads/PBL_2012_From-combating-to-managing demographic-decline-in-the-netherlands.pdf

Ratcliffe, A. (2015). Wealth effects, local area attributes, and economic prospects: On the relationship between house prices and mental wellbeing. Review of Income and Wealth, 61(1), 75-92.

Regio Groningen-Assen (2019). Woningmarktmonitor 2-2019. Groningen.

Rijnks, R. (2020). Subjective Well-Being in a Spatial Context. University of Groningen. https://doi.org/10.33612/diss.133465113

Shotton-Pugh, A. (2019, September 20). *Are Brits Happy With Their Homes?* Main Blog | Terrys Fabrics. https://www.terrysfabrics.co.uk/blogs/main-blog/are-brits-happy-with-their-homes

Sociaal Planbureau Groningen. (2019). Woningvoorraad krimpgebieden Groningen concentreert zich langzaam in grote kernen. Available at: https://sociaalplanbureaugroningen.nl/woningvoorraad-krimpgebieden-groningen-concentreert-zch-langzaam-grote-kernen/ (Accessed: 26 November 2020)

Speare, A. (1974). Residential Satisfaction as an Intervening Variable in Residential Mobility. *Demography*, *11*(2), 173. https://doi.org/10.2307/2060556

Statistics Netherlands. (2014, May 22). *2.5 percent of Dutch housing stock vacant for extended period of time*.https://www.cbs.nl/en-gb/news/2014/21/2-5-percent-of-dutch-housing-stock-vacant-for extended-period-of-time

Voigt, P., & Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing.

Wigren, R., & Wilhelmsson, M. (2007). Housing stock and price adjustments in 12 West European countries between 1976 and 1999. Housing, Theory and Society, 24(2), 133-154.

Wolff, M., & Wiechmann, T. (2014). Indicators to measure shrinking cities. Martinez-Fernandez, Cristina; Fol, Sylvie; Weyman, Tamara; Musterd, Sako: A Conceptual Framework for Shrinking Cities. COST Action TU0803: Cities Re-growing Smaller (CIRES).

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., ... & Narayanan, A. (2017). Ten simple rules for responsible big data research.

VIII. Appendix I - Progress of the dashboard

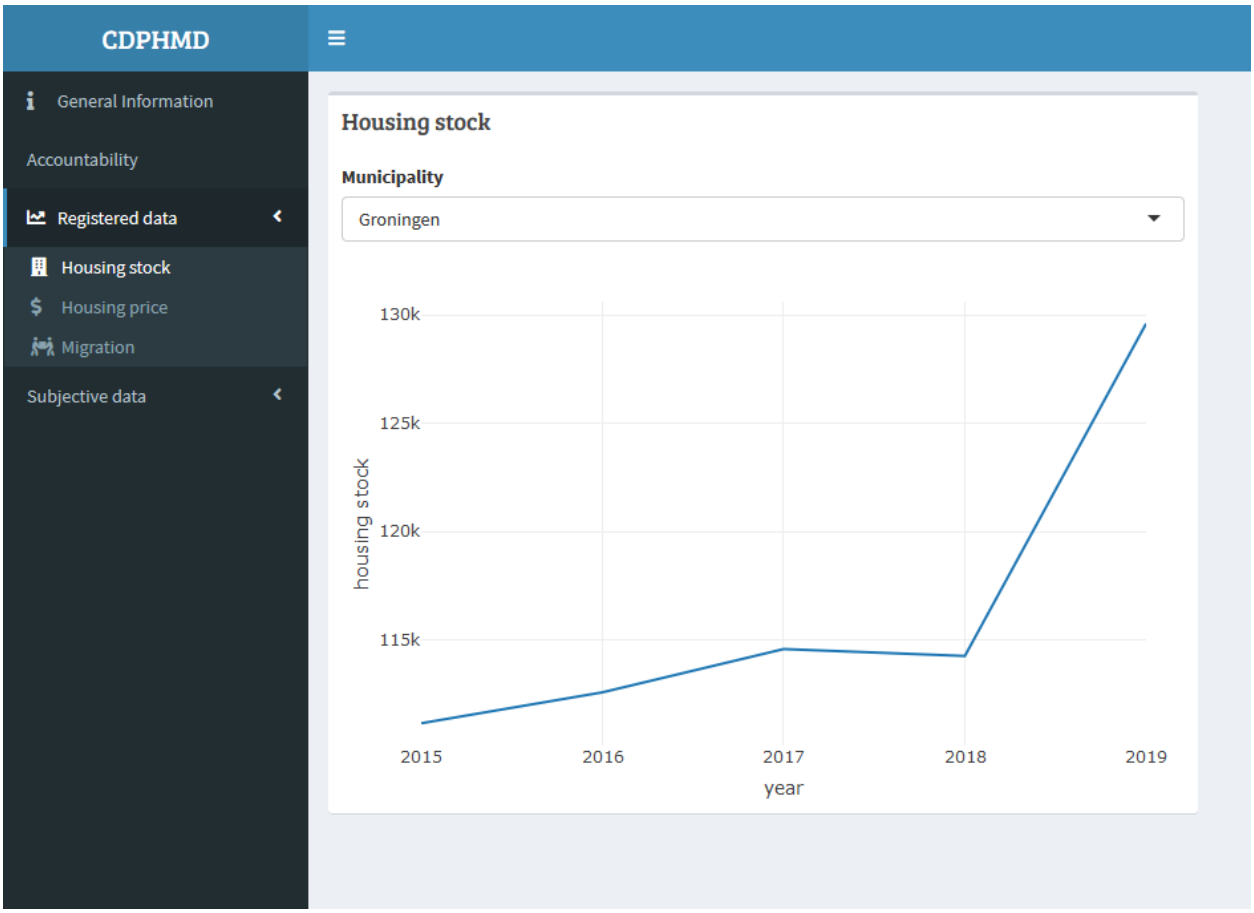

Figure 1 - Screenshot of housing stock page in the dashboard. It shows a line graph of the total housing stock for the municipality of Groningen over the period 2015-2019.

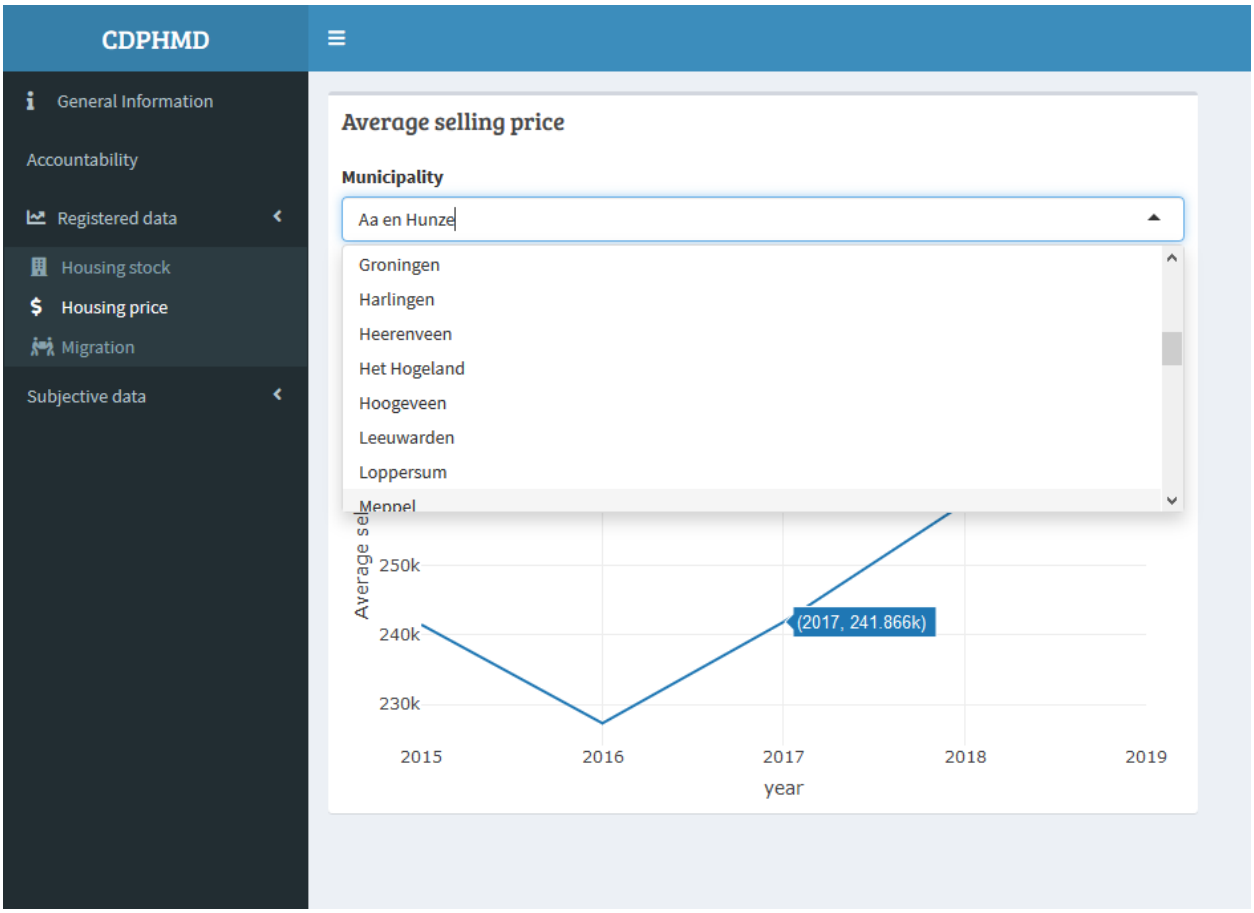

Figure 2 - Screenshot of the dashboard showing interactive functionalities. The dropdown box lets the user select which municipality to (in this case) view the average selling price of houses for. Hovering the mouse over the line shows a small blue box containing the exact coordinates of the closest data point.