Introduction to Data Science - Project

Analysis of Estonian drinking water

Mihkel Vaino & Kätlin Protsin

Business understanding

The idea behind this project was born from this year's events happening in Estonia when in May 2023 there were news that the drinking water in Kuressaare had been contaminated. Due to this many people were coming down with diarrhea, vomiting and abdominal pain which sent them to hospital. After investigation it turned out that the drinking water in Kuressaare area had been contaminated with Escheria coli bacteria and it took multiple days to fix the issue.

Although the quality of drinking water in Estonia has generally been quite high thanks to regulations and frameworks, the above described event shows that keeping an eye on the quality of drinking water is important to ensure safe and healthy drinking water to all Estonians.

The quality of water largely depends on the hydrogeological conditions of the water formation area, which is why the composition of water varies in different groundwater layer and regions. Water quality can be compromised by deteriorated pipelines and tanks, insufficient water movement in pipes, frequent water interruptions, contaminations etc. The indicators of drinking water quality are divided into three groups: microbiological, chemical, indicators. Microbiological and chemical requirements directly characterize a threat to health. Indicator parameters affect the organoleptic properties of water and indicate the overall pollution of water. Exceeding the limit concentrations of indicator parameters deteriorates the conditions for consumer water use and quality of life, but does not pose a direct threat to health.

One of the goals within this analysis is to give an overview of the quality of water in Estonia in general. This would include investigating different indicators of drinking water quality within different counties of Estonia. The results will be visualised in a map so it would be easier to compare different regions and see how the indicators have changed over time. During this analysis it will also be brought out what have been the main indicators which have caused the drinking water quality to turn bad. We can propose a hypothesis suggesting that the primary cause of the deterioration in drinking water quality is the elevated presence of coliform bacteria.

Another goal with this study is to examine the trends of drinking water quality over time. The main research question here is it see whether the quality of water has gotten worse over time (eg. has the bacteria been cumulating) or are there any other trends visible from the data. There is data regarding drinking water from 2012 which makes it a period of 12 years.

Furthermore there are plans to analyse how the quality of water is in correlation with other environmental factors using pollution load indicator, number of residual contamination sources etc.

Data understanding

The main datasets used in this project are taken from Terviseamet's open data source which are "Joogiveeallikate veeproovid" (drinking water test) and "Veevärkide veeproovid" (plumbing test). Data is given in xml format, which contains nested data with 0-n subvalues of similar type and therefore has to be formatted into a flattened shape. Furthermore data contains unnecessary values that do not contribute towards our research that should be removed (eg. the person who conducted the test, as it is presumed the testers followed a standard procedure and did not alter the results), fields that we are interested in are: water source (also test location), time, tests for different chemicals and if they are within regulated norms (specifics tested have some variation from test to test, but we should have enough test data nevertheless). There are around 200 drinking water tests and 4000 plumbing tests per year. Also we will use table "Pinnaveekogudesse juhitud heitvee reostuskoormus maakonna järgi" from Statistikaamet which can be downloaded already in csv format and does not have to be formatted before importing. The wastewater table is not as specific as the previous table, but should give us enough overview to reach conclusions if there is any correlation between them.

On basic examination in the clean water data there are no null values in type of the water, location and time. Every entry contains at least one test, but the amount and type of tests vary with regular checks being more thorough compared to self-ordered tests. The tests without a specific type of test will be left as null and ignored. For now we saw small trend changes in the tests over the years, with them usually becoming more thorough over the years. The clean water tables have differences in quality level classifications with plumbing data having just suitable/unsuitable and drinking water has quality grades 1 through 3 and unsuitable. The drinking water quality level will be formatted into binary values (suitable/unsuitable) as we are more interested in the chemical/bacteria values anyway the classification will be left in the data more as a reference (eg. comparing with news of that time). The wastewater table values seem to vary over the years (2012-2023) so it will be interesting to see if there is any correlation between that and clean water quality.

In initial data exploration the quality of the data seems to be good and consistent, with no apparent gaps that would make analysing it difficult.

Planning your project

- Bringing the data to Python using ElementTree as the xml files contain nested data or option two would be to use the automatic Power BI parser to get the data already in a tabular format. This step also includes parsing xml format into a dataframe and flattening data using recursive algorithms
  We estimate approx. 5 hours for this.

- Getting the data ready to use for analysis. This means cleaning the data (removing unnecessary columns, checking for nulls, unifying data eg. county data, formatting drinking water quality to binary values)
  We estimate approx. 6 hours for this from both team members.
- Investigating the indicators used to measure drinking water quality and analysing if there is any indicator that causes the deterioration of water quality the most.
  Mihkel will focus on this part and we estimate approx. 5h.
- Visualising data - there is a plan to use Tableau or Power BI for visualising county level data.
  As Kätlin has used the tools before we estimate this will take approx. 5h.
- Learning how to use median filtering which was a method suggested as a tool in time series analysis. Using median filtering we plan to investigate the trends in the data over time.
  This task will be done together and we estimate approx. 8h from each.
- Using methods studied during this course to analyse how the quality of water is in correlation with another environmental factor. The specific method will clarify once working with the data.
  This task will be done together and we estimate approx. 10h from each.