# Programming 1: Final assignment.

In this document the choices made during the final assignment are explained.

## The data

All the used data comes from the entirety of the Netherlands. The domain category of the data is: safety research. 3 datasets are used in total excluding metadata. The datasets are:

A perceived safety feeling dataset obtained from the CBS. It comes from a safety survey in which many questions are asked about the respondent’s safety experience in the Netherlands. For the assignment the data is loaded in from the web using a python library provided by the CBS. For the assignment the dataset has been specifically filtered on age groups & years, as well as the types of crimes.

A dataset containing information about how many people have been the victims of crimes. This data was obtained through the same survey as the perceived safety dataset. The dataset is loaded in & filtered the same way as the perceived safety dataset.

A dataset regarding theft registered by the police obtained from the CBS. This dataset contains information about thefts registered by the police per year in absolute numbers. This dataset does not contain information per age group and is only filtered on years and types of crime. While it is possible to load this dataset in through the web the same way as the other files. The choice has been made to store the file prefiltered in the GitHub repository. This is because the file is large and only a select few cells are necessary for the assignment.

Links to the files can be found at the end of this report.

## The research question

The initial research question was looking for correlation between the perceived safety and actual safety in different age groups. Sadly, because the two files with age groups are coming from the same survey with the same respondents there was always going to be an inherent bias and correlation. Therefor a third and independent dataset was brought in, only this dataset did not contain a separation in data for age groups.

There are two thing that the 3 datasets had in common, these two things are that the datasets are separated per year, and that they have data about the same subjects: Total crime, pickpocketing, robbery, and burglary.

Based on these factors a new research question was formed regarding the trends of perceived safety and the committed crimes reported by the police.

When the thefts reported by the police change, does the perceived safety follow in the same direction?

## The data analysis approach

Most of the data transforming is done within a function, this choice was made because it makes it easy to apply the function to multiple years of data without having to repeat any code. It also future proofs the code as when the CBS adds new data, a new year can simply be added to the year list to include the new data. The police reported dataset is less future proof as it is not loaded in through an API and therefore will have to be redownloaded.

The most important thing for the data analysis was to match the crime types of the 3 datasets together so that they could be compared. To do this the preview table of the datasets on cbs.nl was used as this table provides information about what the different columns exactly mean. In the notebook a small summary about what each group means and contains is given.

After filtering out and obtaining the right groups another important step was normalizing the data. The police theft file contains absolute numbers, and the survey files contains *‘% of people’* therefor they could at first not be compared and had to be normalized. The chosen normalisation function is a vector norm normalization that scales the values from 0 to 1.

## The data visualization approach

For investigating the correlation between the two survey datasets per age group 3 different graphs were made:

A grouped bar chart plotting the % of unsafety against the % of victims for each age group per a certain group. This plot clearly shows the differences between age groups. For most categories and years, young people feel unsafe more often, but are also the victims of crime more often.

A heatmap showing correlation between either all the fields or only the grouped pairs was also created. For most years and groups there is a strong correlation between the % of unsafety and the % victims of that type of crime. This could also be seen from the bar graph when unsafety goes down or up the reported total crime follows. Only in 2019 the burglary group has a slight inverse correlation with each other. This inverse correlation can also be seen in the bar graph.

A barbell plot can also be found on this page. This plot presents the same information as the bar chart, just in a different style.

**Yearly trends**

To see the yearly trends between the 3 different datasets, 3 line graphs were made. The first graph is the most important one in helping to answer the research question.

The first line graph shows the normalized values of the different groups plotted from 2012 to 2019. (Skipping 2018) In this graph it is clearly visible that the ‘Total’ group, the pickpocketing group, and burglary group all behave in the same way and are on a downward trend from 2012.

The other two line graphs show the trends of the non normalized values of the 2 survey datasets and the police reported dataset. These graphs show the same downward trends as the first line graph but this time with representative data instead of normalized values.

All graphs were made with the plotly library, as it is an easy to use package which allows for clear customizable visualisations.

## Link to files

Safety survey regarding perception of safety:

<https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=81881NED&_theme=407>

Safety survey regarding victims of crimes:

<https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=83095NED&_theme=409>

Police registered theft dataset: <https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=83651NED&_theme=401>

**Packages used**

For the loading in of the data the [Cbsodata](https://pypi.org/project/cbsodata/) & [Pandas](https://pandas.pydata.org/) packages were used.

For the analysis and transforming of the data, [Pandas](https://pandas.pydata.org/), [Numpy](https://numpy.org/) and [sklearn](https://scikit-learn.org/stable/) were used.

For plotting and creating the dashboard, [plotly express](https://plotly.com/python/plotly-express/) & [panel](https://panel.holoviz.org/) were used.