**Programming 1**

Main question was 'Does the land use of a province (in the Netherlands) affect the number of people who die from a certain disease?' this is more specified during the investigation to:

Does the land use (Total Traffic Area, TotalBuiltTerrein, TotaalRecreatieterrein, TotaalAgrarischTerrein, TotaalBosEnOpenNaturalTerrein) of a province (in the Netherlands) affect the number of people who die from Diseases of Respiratory System?

To answer the above question, the following datasets were used:

* **Land use (.csv):** <https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=70262ned&_theme=298>   
  The purpose of this table is to provide insight into the use of the available space in the Netherlands and the changes taking place therein.

Only the rows containing provinces were included in this study. A selection has also been made on the columns of land use.

* **Deceased (.csv):** <https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=80142ned&_theme=289>   
  This table contains figures on deaths by underlying cause of death by region. The causes of death are divided into four main groups: neoplasms, diseases of the cardiovascular system, diseases of the respiratory system and external causes (non-natural causes of death). The figures are shown according to the municipality of registration. The regions are composed of municipalities. In addition to municipalities, data is available for the following regions: province, part of the country and the whole of the Netherlands.

Only the rows containing provinces were included in this study.

**Population development (.csv):** <https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=37259ned&_theme=257>   
Population development in the Netherlands through births, deaths and migration by gender and region.  
Only the rows containing provinces were included in this study.

* **Provinces borders (.geojson):**  
  <https://www.webuildinternet.com/articles/2015-07-19-geojson-data-of-the-netherlands/provinces.geojson>   
  This contains the border coordinates of the provinces of the Netherlands.

The first three datasets are fetched with an api, the script of which is on github.

The top 3 datasets are stored in a pandas dataframe. The last dataset was loaded as a json file and is partly stored in a pandas data frame, but also retrieved as a dictionary/json. In the last dataset, more use was made of the dictionary because it worked more easily with these data.

The first three datasets have been filtered by doing the following things on them:

* Columns are selected or dropped
* Cnly the rows of provinces are selected
* Percentages are calculated
  + deceased persons compared to persons on December 31 of that year in die provincie
  + land use in relation to the total area of that province in that year
* Data frames are merged

Values have been included with the last dataset, in order to be able to add a hover to the map (for which this data is only used).

After filtering, the data (of the first three datasets) was checked for data types, missing values and counts of provinces. Subsequently, many plots were made to visualize the data. During the first analysis, all columns for causes of death and a selected number of columns for land use were included (fig 1 – 10).  
Later a smaller selection was made by focusing on Diseases of the Respiratory System (causes of death) and the following columns of land use: Total Traffic Area, Total Built Area, Total Recreation area, Total Agricultural Site and Total Forest And Open Natural Terrain (fig 11-17).  
The last plots were only created between Diseases of the Respiratory System and Total Forest And Open Natural Terrain (fig 18-20).

Overview and an explanation of why these plots were chosen:

* figure 1-4: **Stacked and grouped barplot**  
  In these plots you can easily compare the years per provinces or provinces per year between the percentages of deceased and land use.
* figure 5: **Heatmap**  
  In a heatmap you can easily see whether there are columns that correlate with each other.
* figure 6-8: **Boxplots**  
  In a box plot you can quickly see that the variance is equal between different groups. In this case periods and regions.
* figure 9: **QQ-plots**  
  In a QQ plot one can see the data is normally distributed. The data is normally distributed if it falls within the 95% confidence internal.
* figure 10-11: **Histogram (normal distribution)**  
  The histograms are made to also see if the data is normally distributed. But then with a different visualization.
* figure 12-15: **Scatterplot**  
  The scatter plots are made to see if there is a correlation between a column of land use and a column of deceased
* figure 16: **Map**  
  The map is made to view the values of a column of land use and a column of deceased in 1 storage on a map. You can easily compare these with each other. There is also a hover, so you can see the values of that specific column per province.
* figure 17-20: **OLS**  
  \* consists of: Result summary, seaborn OLS, bokeh OLS.ordinary least squares (OLS) is a linear least square method type for estimating the unknown parameters in a linear regression model.  
  HC1 (heteroscedasticity-consistent 1) was used for the ruboost OLS.  
  HC1 (MacKinnon and White, 1985) is the most commonly used robust standard error estimator.

The plots and conclusions of the plots can be found in the document 'Final assignment Anne van Ewijk.ipynb'.

**Conclusion**

The main question was 'Does the land use of a province (in the Netherlands) affect the number of people who die from a certain disease?' this is more specified during the investigation to:

Does the land use (Total Traffic Area, TotalBuiltTerrein, TotaalRecreatieterrein, TotaalAgrarischTerrein, TotaalBosEnOpenNaturalTerrein) of a province (in the Netherlands) affect the number of people who die from Diseases of Respiratory System?

The plots show, among other things, that there is no correlation between these columns and that the data is not normally distributed.  
Furthermore, the following comes from the scatter plots ......  
The OLS test shows that Total Traffic Area, Total Agricultural Area, Total Forest and Open Natural Area are significant, but these have such a small slope that this may be negligible.

More research would need to be done to reach a final conclusion.

In follow-up research, the following things should be carefully considered:

* You must make sure that there is nothing that correlates with both land use and deceased that you do not use in your model. Otherwise, you don't measure the impact of one on the other, but just how much they roughly match. An example could be that, for example, older people live in Zeeland and that this is the result of more deaths and not, for example, the percentage of traffic areas. So if you check / correct for these things you can see what the impast of land use or death is.
* You cannot say from an academic perspective that something has an impact on something else with just these two variables.