# CW01 - Analysing UK Health Profiles

**OVERVIEW of TASK:**

In this coursework, your task is to inform the health authorities on the relations between health indicators and socio-economic and demographic indicators. You will need to structure an analysis strategy, process and merge the data, investigate the relations, model a particular relation and inform the authorities on possible future outcomes. A detailed description of your tasks in this coursework are below. We continue now with a description of the data you will be working with.

**DATA:**

In this case, we are lucky that the authorities have nicely structured the data and made sure that it is easily possible to join the different data sets. In this coursework, we will be using the [information collected in 2013](http://www.apho.org.uk/resource/view.aspx?RID=126453). The data is for each local authority in England and presented in the form of indicators that have been carefully processed and made available for analysis.. Within this coursework we are interested in the 32 indicators each of which relates to a particular aspect and categorised under these groups:

* Our communities
* Children's and young people's health
* Adults' health and lifestyle
* Disease and poor health
* Life expectancy and causes of death

[On the link above](http://www.apho.org.uk/resource/view.aspx?RID=126453), you can download data on an indicator in either CSV or Excel format. Each file starts with a meta-data section where the indicator is described so these brief information should be sufficient to understand the contents of the file. However, in order to get a deeper understanding of the indicators and to read about how they have been collected, you can [have a look at this detailed document about the indicators](http://staff.city.ac.uk/~sbbk529/Teaching/Resources/INM430/CW01/PHEHealthProfilesIndicatorGuide2013.pdf).

In each of the indicator data files (whether xlsx or csv), you will see columns where the data resides. In order to link these data sources to each other, we will use the column named "ONS Code (new)" and the actual indicator values are under the column "Indicator value". These are the two columns that you need from each indicator data table. (Optional) If you would like to read more about how the indicators have been generated, you can have [a look at this file](http://www.apho.org.uk/resource/view.aspx?RID=44584).

**Tasks:**

As also described above, the expectation of the health authorities is to learn about the relation between the health indicators and societal information. In order to help them in this tasks, you need to carry out the following sub-tasks and report about them:

**Task 1 -- Build an Analysis Strategy**

Develop an analysis strategy and write a brief description about your analysis outline. This should contain your comments on:

* Identify the dependent and independent indicator categories
* Choose one dependent variable and at least two independent variables
* Describe what steps you are taking during the analysis

**Task 1**

Overview - Analysis Strategy

The purpose of this analysis strategy is to clearly outline a logical process such that the initial requirements from UK authorities that is to better understand the data that has been collected. Through the use of the exploratory analysis the aim is to understand and produce new insights, the steps of this strategy have been summarised below:

* Collection
* Preparation of the Data for Analysis
* Exploratory Data Analysis
* Development of Models and Visualisation

*Collection*

The collection of the data has already been completed by the local authorities and is stored remotely on a web server. The data is in the form of thirty two individual files (CSV and XLSX versions exist). The first process will be to merge all of the datasets together, this will provide an easier means of comparison of the data, as it will be in one data frame.

*Preparation of Data for Analysis*

Once the data has been merged into a Master data frame that includes the name of the Area, ONS codes and each of the thirty two indicators the preparation can begin. Inspection that the data has been merged accurately is essential before continuing, this can also be confirmed by running simple summary steps as confirmation. Identification of any missing values at this stage is key as this may have a profound affect on the on any transformation that may need to be done or could lead to spurious predictions. A decision also of how to deal with them needs to be decided at this stage. Finally, any transformations of the data need to be applied, this may be to normalise the data or alternately the use of dimension reduction (PCA) to consolidated the variables. This can be achieved through the inspection of the descriptive statistics complemented by visuals to support decisions on how to deal with the data.

*Exploratory Data Analysis*

At this stage the data is in an appropriate structure of analysis, data mining techniques to identify outliers, understanding the correlations between variables and basic statistical tests can be initial processed on the data. These tests could include testing for normality, correlation (Spearman, Pearson) and t-tests for significance of the variables. Iteratively moving between the previous stage and this stage to transform the data depending on the output of the analysis at this stage.

*Development of Models and Visualisation*

Once Data Analysis has been completed, the development of statistical models can now be produced. This is dependent on the analysis up till this point, where the goal to inform authorities of insights into the data, through the use of statistical models. Once the model have been generated, the correct use of visualisation needs to be obtained to correctly to convey the message that is sought to be shown.

The table below describes the identity of each of the categories:



The dependent categories are dependent of a variety of measures namely societal status and services available in each of authorities. Whereas the independent categories have no immediate dependencies.

Dependent variable:

1. Deprivation

Independent variables:

1. Drug Misuse
2. Acute sexually transmitted infections
3. Obese Children (Year 6)

**Task 2 -- Prepare data Script DONE**

Although the data comes highly structured and processed, there are still steps needed to be taken to get it ready for analysis. You may need to, load the data, merge and normalise accordingly. Describe the steps you have taken, talk about any problems that might occur and justify any choices you've made in this step, e.g., why you chose a particular normalisation method?

These operations needs to be done with Python scripts similar to what we've done during the lab sessions.

**Task 3 -- Perform analysis**

Once the data is ready, perform an analysis that will lead to a better understanding of the data. A couple of analytical insights that can be reported here are:

* Have you come across any outliers / anomalies?
* What are the relations between the dependent and independent indicators?
* Can we build a model to characterise these relations?

Your findings should inform the authorities about any relations, anomalies on the indicators you've chosen. You need to comment on the certainty of the findings and evaluate the results you've obtained. Feel free to support your findings with plots you might generate.

**Submitting Coursework:**

* Report : Submit a report that describes the observations and finding in each step. Clearly separate the input into the three tasks in your report. Your report should cover all the important details but should not be longer than 2-3 pages.
* Python script: Submit a Python script with codes that you've written to carry out the task.Clearly mark the parts where you do the analysis for the separete tasks. Remember to comment steps briefly. Here is a template [how this file might look like](http://staff.city.ac.uk/~sbbk529/Teaching/Resources/INM430/CW01/cw01_template.py).

Put these two files in a zip package and name the zip file as name\_surname.zip and submit it through the submission page which will be setup and listed on the module's page.

The submission deadline is 12pm on November 17th before our lecture that day. Please be punctual and do not leave your submission to the very last minute.