

# SIG731 2023: Task 7HD

## Data Mining Challenge

Last updated: 2023-11-24

### Contents

<b>1 Task</b>	<b>1</b>
<b>2 Artefacts</b>	<b>1</b>

Tasks 5–8 are not obligatory; you can submit them in any order (or decide not to tackle them at all). C/D/HD is merely a subjective estimate of their difficulty level. For each task that you successfully complete, you score 10 points (and for those that are not 100% correct, no points will be given).

### 1 Task

Create a single Jupyter/IPython notebook (see the *Artefacts* section below for all the requirements), where you perform what follows.

1. Download at least five different datasets that are part of the NHANES study; see <https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2021> and <https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020>. Merge them into a single data frame.
2. Using the `bokeh` package, which you will have to learn yourself (this is part of this HD-level task), create at least five nontrivial *interactive* data visualisations and/or tables.
3. Draw insightful and interesting conclusions. Do not forget to reflect on the potential data privacy and ethics issues that arise during the data analysis process.

Make it aesthetic and interesting to read.

*This HD-level task is purposely under-defined – you will not be told precisely what to do. Your aim is to discover, visualise, and explain some **interesting** relationships between **many** data features.*

In the course of the report preparation, you should apply a wide range of data frame wrangling techniques, including filtering, aggregation in groups, missing value handling, column transformation, etc.

Do not use pie charts (as we discussed during the lecture). Please go beyond simple bar/box plots and histograms. The charts/tables must be interactive; a reader should be able to use slider bars, text boxes, etc. to control the data/view.

For some inspiration, you might want to take a look at some research papers citing the NHANES study: [https://scholar.google.com/scholar?hl=en&as\\_sdt=0%2C5&q=nhanes](https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=nhanes).

## 2 Artefacts

Make sure that your notebook has a **readable structure**; in particular, that it is divided into sections. Use rich Markdown formatting (text in dedicated Markdown chunks – not just Python comments).

Do not include the questions/tasks from the task specification. Your notebook should read nicely and smoothly – like a report from data analysis that you designed yourself. Make the flow read natural (e.g., *First, let us load the data on... Then, let us determine... etc.*). Imagine it is a piece of work that you would like to show to your manager or clients — you certainly want to make a good impression. Check your spelling and grammar. Also, use formal language.

At the start of the notebook, you need to provide: the **title** of the report (e.g., *Task 42: How Much I Love This Unit*), your **name**, **student number**, and **email address**.

Then, add 1–2 introductory paragraphs (an introduction/abstract – what the task is about).

Before each nontrivial code chunk, briefly **explain** what its purpose is. After each code chunk, **summarise and discuss the obtained results** (in a few sentences).

Conclude the report with 1–2 paragraphs (summary/discussion/possible extensions of the analysis etc.).

---

### Checklist:

1. Header, introduction, conclusion (Markdown chunks).
2. Text divided into sections, all major code chunks commented and discussed in your own words (Markdown chunks).
3. Every subtask addressed/solved. In particular, all reference results that are part of the task specification have been reproduced (plots, computed aggregates, etc.).
4. The report is readable and neat. In particular:
  - all code lines are visible in their entirety (they are not too long),
  - code chunks use consecutive numbering (select *Kernel - Restart and Run All* from the Jupyter menu),
  - rich Markdown formatting is used (# Section Title, \* bullet list, 1. enumerated list, | table |, *\*italic\**, etc.),
  - the printing of unnecessary/intermediate objects is minimised (focus on reporting the results specifically requested in the task specification).

Submissions which do not *fully* (100%) conform to the task specification *on* the cut-off date will be marked as FAIL.

Good luck!