SIT220/731 2025.T1: Task 7HD

Data Mining Challenge

Last updated: 2025-01-30

Contents

1 Introduction

> 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).

This task is due on **Week 11 (Friday)**. Start tackling this task as early as possible, because at least 2–3 iterations are usually needed before you get on the right track. In case of any problems/questions, do hot hesitate to attend our on-campus/online classes.

Submitting after the aforementioned due date might incur a late penalty. The **cut-off date is Week 12 (Friday)**. There will be **no extensions** and no solutions will be accepted thereafter. At that time, if your submission is not 100% complete, it will be marked as FAIL, without the possibility of correcting and resubmitting. To ensure a fair environment for all, we are always very strict about deadlines.

2 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?Cycle=2021-2025 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.

3 Additional Tasks for Postgraduate (SIT731) Students (*)

There are no specific additional tasks, because the whole exercise has an open-ended formulation.

4 Artefacts

The solution to the task must be included in a single Jupyter/IPython notebook (an .ipynb file) running against a Python 3 kernel. The use of G**gle Colab is discouraged. Nothing beats a locally-installed version where you have full control over the environment. Do not become dependent on third-party middlemen/distributors. Choose freedom instead.

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**, **email address**, and whether you are an **undergraduate (SIT220) or postgraduate (SIT731)** student.

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.).

Limitations of the OnTrack ipynb-to-pdf renderer:

Ensure that your report as seen in OnTrack is aesthetic (see *Download submision PDF* after uploading the .ipynb file). The OnTrack ipynb-to-pdf renderer is imperfect. We work with what we have. Here are the most common Markdown-related errors.

- Do not include any externally loaded images (via the `![label](href)` Markdown command), for they lead to upload errors.

- Do not input HTML code in Markdown.

- Make sure you leave one blank line before and after each paragraph and bullet list. Do not use backslashes at the end of the line.

- Currently, also *LaTeXformulae* and Markdown tables are not recognised. However, they do not lead to any errors.

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!

5 Intended Learning Outcomes

ULO	Is Related?
ULO1 (Data Processing/Wrangling)	YES
ULO2 (Data Discovery/Extraction)	YES
ULO3 (Requirement Analysis/Data Sources)	YES
ULO4 (Exploratory Data Analysis)	YES
ULO5 (Data Privacy and Ethics)	YES