

ASSIGNMENT

Data Science – Data Collection and Data cleaning

You are asked to create Python programs to process the file kaggle\_survey.zip which includes the answers from the respondents who participated 2019-2021 in the very famous Data Science Kaggle Survey. Use **seaborn** (Other visualization liberalises are not allowed, except for using Matplotlib to specify figures’ size) and **pandas** to support the data processing and analysis.

In addition to the codes for Problem 1-5, you also need to submit a concise report (no more than 5 pages, no including references) to explain how to solve Problem 2-5. You can mention your codes with the cell/line numbers but do not include raw codes. Please comment your code properly.

**Total: 100%**

**Problem 1. (10%)**

Extract all the files from Kaggle\_Survey.zip **using Python code** into a folder named “Kaggle”. Make sure none of the files inside “Kaggle” folder is a zip file.

**Problem 2. (30%)**

Study the extracted files and observe the questions that appears in all the three surveys. Then write a python program to create a new csv call “Kaggle\_survey 2019-2021.csv” to save the corresponding data of those questions, also add a new column “year of the answer”. Note: some questions may slightly change the expression or the options. Explain in you report how you deal with the **data merging**.

**Problem 3. (30%)**

Use the **data cleaning** methods you learned in the lectures to clean and process the data “Kaggle\_survey 2019-2021.csv” and save to another csv called “Kaggle\_survey 2019-2021\_cleaned.csv”.

**Problem 4.** **(10%)**

Based on the data in “Kaggle\_survey 2019-2021\_cleaned.csv”, investigate the top 5 programming languages (5%) and the top 5 visualization libraries/ tools (5%) used by the senior (more than 5-year programming experience) Data Scientists. Display the results of 2019,2020 and 2021 separately in visual graphs and make your conclusion.

**Problem 5.** **(20%)**

Based on the data in “Kaggle\_survey 2019-2021\_cleaned.csv”, explain the world-wide situation of Woman in Data Science with the support of visual graphs.

**Plagiarism:** You must not plagiarise your work. Attempts to hide plagiarism by simply changing comments/variable names will be detected. You should have been made aware of the Durham University policy on plagiarism.

To submit your work, create a directory named as your username (e.g. cxfh123). Place all required files in this directory using, ZIP compress/archive this entire directory structure (not .rar or .z7 or anything else - as this breaks the automated extract/test tools).