

**Background**

One of the largest food companies in the world has engaged Streetbees to conduct Life Moments research to understand what people eat. Rather than relying on people's memory of what they eat, Streetbees has asked participants to log every meal they have for a full week by taking photos and telling us what they are eating at the moment.

**Case Study**

As part of this project we have collected new data but we are also using submissions from a previous project that we had collected for this client. For each project we have a background file that contains the background profiling questions and the logs that capture the relevant questions captured during in-the-moment food consumption. *project\_1\_background.csv* and *project\_1\_logs.csv* contains the details for the previous project whereas *project\_2\_background.csv* and *project\_2\_logs.csv* contains the more recent data we have collected.

Throughout this task we’re looking for production grade code, thinking about readability, reusability and performance.

**Task 1 – Identification of issues and improvements**

Create a data pipeline that combines these datasets and performs a number of data cleaning tasks. The required operations are outlined below:

* The 2 logs and the 2 backgrounds need to be merged - column names need to be mapped for it.
* User genders are spelt / capitalised differently
* Some location names are codes, others are full names. Need to be the same.
* Mismatched user age groups.
* Duplicate columns in the same dataset which need to be merged
* *Level2dish\_coded* column is a dictionary instead of a string
* *Questions\_135633\_and\_who\_are\_you\_sharing\_your\_home\_with* - {["Children over 18 years old"]} - turn to normal string (e.g. ["Children over 18 years old"])
* Some columns have duplicate values which need to be merged (e.g. "Kids/ family like it", "Kids / family like it")

Deliverables:

* Python file with the above functions

**Task 2 – Create a package**

Now that the functions are created they need to be made available to be used by other teams. Please create a package from where the other teams could pull them and use them in their scripts. Feel free to modify them to make them more generic so datasets with similar, but not identical issues, can be cleaned too. Don’t forget to add unit tests and documentation to them as well - we don’t expect 100% test coverage, more just a few examples that demonstrate unit test experience.

Deliverables:

* Zip folder with the code. Note: please don’t make any repos or data public on GitHub

**Task 3 – Demonstration of the package**

Now that the package is created, please create a Python script or a Jupyter notebook where you import your package and use its functions to clean and prepare the attached datasets.

Deliverables:

* Python script or a Jupyter Notebook with an example for the given dataset
* Brief description on how to run it