Data Challenge

**Context**

A company has collected data (ie. different tables) from different collaborators and now needs to transform the data similarly to UK Biobank in order to run joint analysis

**UK Biobank format**

Imagine the following raw data

|  |  |  |
| --- | --- | --- |
| participant\_id | BMI\_visit1 | BMI\_visit2 |
| 1101010 | 10 | 20 |
| 1101010 | 14 | 13 |
| 1101010 | 12 | NA |
| 1101013 | 35 | 32 |
| 1101013 | 27 | NA |

In here the same person (1101010) measured BMI three times for the first visit (10,14,12) and two times for the second visit (20, 13)

In UK Biobank format the data would be wide (ie. each row uniquely identifies a participant) and the column names would have the following convention:

**PhenotypeID-Instance.Array**

* PhenotypeID is a unique id to assign to the phenotype (888 in the example)
* Instance is the number of the time point (two time points in the example)
* Array is the number of the repeated measure (3 measures for participant 1101010)

The final UK Biobank format is

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| participant\_id | 888-0.0 | 888-0.1 | 888-0.2 | 888-1.0 | 888-1.1 | 888-1.2 |
| 1101010 | 10 | 14 | 12 | 20 | 13 | 13 |
| 1101013 | 35 | 27 |  | 32 | NA | NA |

**Challenge:**

The challenge is to transform the data (long format) into UK Biobank format (wide format)

**Acceptance criteria**

* Assign to each unique phenotype a unique ID (eg. “888”)
* For this exercise we do not have time points and all instances can be set to 0
* Final data has to be wide ie. each row identifies a unique participant
* Each column name has to be unique
* Merge tables by participant\_id
* Missing data coded as NA
* Drop columns which are all empty
* Coded phenotypes (found in *data\_dictionary.csv*) have to be transformed. eg. a participant with “fathers\_ethnic\_category” with value “A”, the value has to be transformed to “White: British”
* Any output format is accepted (eg. tsv, csv)
* Any language is accepted (R, Python, …)

Extra criteria (not mandatory)

* Dates in format DD-MM-YYYY
* For dates store hours:minutes into a separate variable called [dateColumnName]\_time