**Data:**

File *cell-count.csv* contains cell count information for various immune cell populations of each patient sample. There are five populations: *b\_cell*, *cd8\_t\_cell*, *cd4\_t\_cell*, *nk\_cell*, and *monocyte*. Each row in the file corresponds to a biological sample.

**Python:**

1. **Please write a python program to convert cell count in *cell-count.csv* to relative frequency (in percentage) of total cell count for each sample.** Total cell count of each sample is the sum of cells in the five populations of that sample. Please return an output file in csv format with cell count and relative frequency of each population of each sample per line. The output file should have the following columns:

***sample***: the sample id as in column *sample* in *cell-count.csv*

***total\_count***: total cell count of *sample*

***population***: name of the immune cell population (e.g. *b\_cell, cd8\_t\_cell, etc.*)

***count***: cell count

***percentage***: relative frequency in percentage

1. Among patients who have treatment *tr1*, we are interested in comparing the differences in cell population relative frequencies of melanoma patients who respond (responders) to *tr1* versus those who do not (non-responders), with the overarching aim of predicting response to treatment *tr1*. Response information can be found in column *response*, with value *y* for responding and value *n* for non-responding. Please only include PBMC (blood) samples.
   1. **For each immune cell population, please generate a boxplot of the population relative frequencies comparing responders versus non-responders.**

**A diagram of a number of cells

Description automatically generated**

* 1. **Which cell populations show a difference between responders and non-responders? Please include statistics to support your conclusion.**

Cell populations showing significant difference between responders and non-responders: ['cd4\_t\_cell', 'monocyte']

Please return both the code and the outputs. Please also specify any dependencies that you use and instructions on how to run your code to reproduce the outputs.

**Database:**

1. **How would you design a database to capture the type of information and data in *cell-count.csv*?** Imagine that you’d have hundreds of projects, thousands of samples and various types of analytics you’d want to perform, including the example analysis of responders versus non-responders comparisons above. Please provide a rough prototype schema.

1. Project Table:

* project\_id (Primary key)
* project\_name
* project\_description
* no\_of\_subjects

2. Subject Table:

* subject\_id (Primary key)
* age
* sex
* condition

3. Treatment Table:

* treatment\_id (Primary key)
* treatment\_name
* treatment\_description

4. Sample Table:

* sample\_id (Primary Key)
* sample\_type
* time\_from\_treatment\_start

5. Project\_sample\_analytics Table:

* project\_sample\_analytics\_id (Primary key)
* project\_id (Foreign key)
* subject\_id (Foreign key)
* sample\_id (Foreign key)
* treatment\_id (Foreign key)
* response
* b\_cell\_counts
* cd8\_t\_cell\_counts
* cd4\_t\_cell\_counts
* nk\_cell\_counts
* monocyte\_counts

1. **What would be some advantages in capturing these information in a database?**

Using database to capture information allows storing the data in an organized and structured manner. This enables data to be retrieved efficiently using queries to perform further analysis to gain insights into complex datasets. It also provides scalability to projects by enabling large amounts of data to be handled efficiently and securely. It also promotes collaboration on projects within and between teams.

1. Based on the schema you provide in (1), **please write a query to summarize the number of subjects available for each condition.**

SELECT condition,  
        *Count*(DISTINCT subject\_id) AS subject\_count  
FROM   Subject  
GROUP  BY condition;

1. **Please write a query that returns all melanoma PBMC samples at baseline (*time\_from\_treatment\_start* is 0) from patients who have treatment *tr1*.**

SELECT psa.\*

FROM project\_sample\_analytics psa

JOIN sample s

ON psa.sample\_id = s.sample\_id

JOIN subject su

ON psa.subject\_id = su.subject\_id

JOIN treatment t

ON psa.treatment\_id = t.treatment\_id

JOIN project p

ON psa.project\_id = p.project\_id

WHERE s.sample\_type = 'PBMC'

AND su.condition = ‘melanoma’

AND s.time\_from\_treatment\_start = 0

AND t.treatment\_name = 'tr1';

1. **Please write queries to provide these following further breakdowns for the samples in (4):** 
   1. How many samples from each project

SELECT p.project\_name,

COUNT(\*) AS sample\_count

FROM Project p

JOIN Project\_sample\_analytics psa

ON p.project\_id = psa.project\_id

JOIN Subject su

ON psa.subject\_id = su.subject\_id

JOIN Sample s

ON psa.sample\_id = s.sample\_id

JOIN Treatment t

ON psa.treatment\_id = t.treatment\_id

WHERE s.sample\_type = 'PBMC'

AND su.condition = ‘melanoma’

AND s.time\_from\_treatment\_start = 0

AND t.treatment\_name = 'tr1'

GROUP BY p.project\_name;

* 1. How many responders/non-responders

SELECT response,

COUNT(\*) AS response\_count

FROM Project\_sample\_analytics psa

JOIN Subject su

ON psa.subject\_id = su.subject\_id

JOIN Sample s

ON psa.sample\_id = s.sample\_id

JOIN Treatment t

ON psa.treatment\_id = t.treatment\_id

WHERE s.sample\_type = 'PBMC'

AND su.condition = ‘melanoma’

AND s.time\_from\_treatment\_start = 0

AND t.treatment\_name = 'tr1'

GROUP BY response;

* 1. How many males, females

SELECT su.sex,

COUNT(\*) AS sex\_count

FROM Subject su

JOIN Project\_sample\_analytics psa

ON psa.subject\_id = su.subject\_id

JOIN Sample s

ON psa.sample\_id = s.sample\_id

JOIN Treatment t

ON psa.treatment\_id = t.treatment\_id

WHERE s.sample\_type = 'PBMC'

AND su.condition = ‘melanoma’

AND s.time\_from\_treatment\_start = 0

AND t.treatment\_name = 'tr1'

GROUP BY su.sex;