

DPRPy 2024/2025

Homework assignment no. 1B (max. = 15 p.)

Maximum grade: 15 p.

Homework should be sent via the **MS Teams** platform. You should send **1 file** containing solutions to tasks, i.e. `Last-name_First-name_HA_1B.R` - an R script **or** `Last-name_First-name_HA_1B.ipynb` - a file prepared with Jupyter Notebook.

Important! If you chose to send R script - output such as execution time comparison, equivalence check be paste into the file as comments. If you chose to send .ipynb notebook each chunk must be evaluated with results visible.

Remember to comment your code and take care of the overall readability of, both, your code and a file itself.

1 Data description

We are working on a simplified dump of anonymised data from the website <https://bicycle.stackexchange.com/>, which consists of the following data frames:

- `Posts.csv`
- `Users.csv`
- `Comments.csv`
- `Votes.csv`
- `Tags.csv`
- `Badges.csv`
- `PostLinks.csv`

Before starting to solve the problems familiarize yourself with the said service and data sets structure (e.g. what information individual columns represent), see

<https://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede/2678#2678>.

2 Tasks description

2.1 Task 1 [5 p.]

Find top 10 users with the highest number of posts marked as accepted answer; include user id, his/her display name, total number of questions, answers and comments created by this user as well his/her overall score.

Prepare two solutions - one where you will use only base (built-in) R functions and one with the **dplyr** package. Compare both the execution times of your solutions (using one call to `microbenchmark::microbenchmark()`) and whether the returned data frames are equivalent to each other (with respect to rows and columns permutation).

2.2 Task 2 [5 p.]

Find location of users that obtained at least one of (Teacher, Explainer, Commentator) badge; sort resulting table according to the total number of posts created from this location by users with those badges and the total number of badges.

Prepare two solutions - one where you will use only base (built-in) R functions and one with the `dplyr` package. Compare, both, the execution times of your solutions (using one call to `microbenchmark::microbenchmark()`) and whether the returned data frames are equivalent to each other (with respect to rows and columns permutation).

3 Task 3 [5 p.]

Reproduce the result of the following query using either R-base functions or `dplyr` package. Check whether the resulting data frame is equivalent to the table obtained by executing the query (see `?sqldf`) with respect to rows and columns permutation. Write a short description what is the result of this query.

```
SELECT
  Posts.Title,
  UpVotesPerYear.Year,
  MAX(UpVotesPerYear.Count) AS Count
FROM (
  SELECT
    PostId,
    COUNT(*) AS Count,
    STRFTIME('%Y', Votes.CreationDate) AS Year
  FROM Votes
  WHERE VoteTypeId=2
  GROUP BY PostId, Year
) AS UpVotesPerYear
JOIN Posts ON Posts.Id=UpVotesPerYear.PostId
WHERE Posts.PostTypeId=1
GROUP BY Year
ORDER BY Year ASC
```