Evaluation Sheet

Group Examination
„Big Data Statistics for R and Python:
Group Examination of Part I"
2019

Team system.time:
- Cedric Müller
- Patrik Münch
- Dominik Wingeier

EVALUATION SUMMARY

Implementation of specific tasks (part I of the examination):
12.5 of 15 points.

Own extension of the research report (part II of the examination):
10 of 10 points.

TOTAL POINTS: 22.5

Feedback

Implementation of specific tasks (part I of the examination)

Task 1: Data gathering (1.5 of 2 points)
The code works like a charm, is clean and well-documented. The team's solution is correctly and efficiently implemented. The exposition of the solution is helpful and correct. The advantages of downloading in batches could be extended regarding the issue of potential network interruptions. Moreover, the approach behind the import of CSVs could be discussed in more detail: which function is used and why? How does it basically work?

Task 2: Data storage and databases (3 of 3 points)
Again, the code works like a charm, is clean and well-documented. The team's solution is correctly and efficiently implemented. In particular, the team had a careful look at how to assign meaningful field types to the newly generated SQL tables. In addition, the team has recognized the trade-off between using more disk space for indices and speeding up queries. Their solution is well-balanced in this regard. The exposition of the solution is clear and correct.

Task 3: Data aggregation (4 of 6 points)
Again, the code works like a charm, is clean and well-documented. The solution to task 1 is not entirely correct. The task is to consider only 'contributions to political committees (other than Super PACs and Hybrid PACs) from an individual, partnership or limited liability company' in this analysis'. These are contributions of type '15' (see https://classic.fec.gov/finance/disclosure/metadata/DataDictionaryTransactionTypeCodes.shtml). In the team's solution type '15' is excluded. Minor point regarding task 2: the loop could be further simplified by just iterating through cycles and selecting/ordering the ranking per cycle (each iteration results in one column, the result is the same). The SQL query in task 3 could be substantially simplified by joining the donations table with the industry-codes table and directly selecting the industry IN ('BUSINESS ASSOCIATIONS', 'PUBLIC SECTOR UNIONS', etc.), see column industry here assets.transparencydata.org.s3.amazonaws.com/docs/catcodes.csv. Minor point: the aggregation could be substantially simplified by relying more on data.table's capabilities than loops (see, for example, dcast()). The exposition is very clear and detailed. The team has done a great job in thinking about a good balance between SQL and in-memory operations in R. The way SQL queries are implemented and explained in the exposition is technically sound. My criticism above simply aims to point out that, given the data and task at hand, there is not really a need for such complex queries in some cases.

Task 4: Visualization (4 of 4 points)
The team has executed the visualization task very well. The code works like a charm and is very clean and well readable. The resulting figure is very clear and professional. The team's exposition is correct, clear, and to the point.

Own extension of the research report (part II of the examination):

Summary/Content and Empirical Approach

The team presents a carefully crafted analysis of an interesting research question. The analysis is interesting, relevant, and well-motivated. The empirical approach is thoroughly explained and justified. Moreover, the team is critically discussing the results.

A further avenue to think about the partly ambiguous results: Could it be that the negative sign of the money coefficient captures some sort of Simpson's paradox in the data? I could imagine that within races, the partial correlation between money and winning probability is often positive (as we would expect) while there might be substantial differences in terms of the narrowness of outcomes and money spent between the races (or more generally, over time). In any event, the team did a great job in this part of the examination.

Code

The code works like a charm. It is very well structured and well documented. The team fulfils the condition of using a virtual memory package as a main aspect of the implementation (here: to compute aggregates).

Presentation of Results

The presentation of the results is very thoughtful, clean, and convincing.