Reproducibility Report for ACM SIGMOD 2020 Paper: "QueryVis: Logic-based Diagrams help Users Understand Complicated SQL Queries Faster"

RAUL CASTRO FERNANDEZ, The University of Chicago, USA

The main results of this paper are supported by the analysis of data collected via a user study. This report focuses on three aspects: i) the user study described in the paper; ii) threats to validity that may have been introduced in the pre-screening/tutorial documents and that could affect the results; iii) inspection of the data analysis code and verifying results are repeatable and that the conclusions reached in the paper follow from the data. We conclude that the results of this paper are supported by the data collected, that the data was collected using appropriate best practices and materials. The detailed analysis conducted by the authors makes the results in the paper stronger.

1 INTRODUCTION

This report summarizes the experience of reproducing the results for the SIGMOD 2020 paper: "'QueryVis: Logic-based Diagrams help Users Understand Complicated SQL Queries Faster"'.

The bulk of the evaluation results in this paper is the analysis of data collected via a user study. Reproducing the results would require: i) collecting data from a different set of participants; ii) repeating the analysis pipeline. Our effort focuses on the latter point. We did not repeat the user study. In addition, this report has also looked into the details of study design, paying attention to potential threats to the validity of the results that would make reproduce the results difficult.

Overall, the procedures followed to conduct the user study are robust, and follow best practices. The analysis is detailed, easy to follow, and logical. The authors provided a materialized version of the results as a PDF document that makes it straightforward to follow their data analysis process. They provide instructions to repeat their analysis.

In summary, the materials increased our confidence in the conclusions reached in the paper.

2 SUBMISSION

The user study conducted in this paper was pre-registered at osf.io, an online portal aimed to help researchers register and document their user studies. This is consistent with best practices. The submission at osf.io contains several important materials, of which I highlight:

- The tutorial that participants had to follow.
- The 6 pre-screening SQL questions.
- The actual user study content received by the users.
- A link to the code used to deploy the study, via Amazon Mechanical Turk.
- PDFs with the contents of the Jupyter notebooks where the bulk data analysis is presented.
- Instructions to set up a python environment that permits repeating the data analysis.

I want to note that the experimental data is also included in the site, and appropriately anonymized, hence respecting participants' privacy and following best practices.

3 HARDWARE AND SOFTWARE ENVIRONMENT

The results can be repeated via a Python environment. The authors provide instructions to set up this environment. I will say the process is straightforward for people with previous experience working with Python virtual environments. A few typical hiccups were found while setting this up. When running on a MacOS, installing Matplotlib is known to throw errors if pkg-config is not installed, for example. Similarly, there's a mismatch between numpy and scipy library versions, due

to the Python version 3.8 instead of 3.6. Although the authors provided a *requirements_basic.txt* which does not indicate versions, there was also a *requirements.txt* which is the de-facto standard file on which the output of pip freeze is stored.

None of these problems were severe, and they can be avoided by capturing the environment completely, for example, via the use of container technology, such as Docker.

4 REPRODUCIBILITY EVALUATION

4.1 Process

To reproduce a user study, one would need to recruit a different set of participants, subject them to the same user study, collect the data, analyze the data following the same procedure followed by the authors, and observe whether one can reach the same conclusions with the same degree of statistical significance[1]. The scope of this document focused instead on 3 aspects:

- Threats to validity. May the tutorial or pre-screening materials may have introduced some kind of bias in the process that would threaten the external validity of the study? and hence, its reproducibility? I studied the paper and materials in detail.
- Were the analysis methods used appropriate and rigorous? I followed the data analysis myself, reasoning from first principles and verifying the code provided by the authors.
- Was it easy to follow the analysis and contrast the results with what reported in the paper?

Overall, I found the process clean, easy to follow, rigorous, and it did increase my confidence in the results presented in the paper. I verified pre-registration timestamps, followed the tutorial, took the pre-screening questions, studied the data analysis results, deployed the Python environment, and was able to repeat the same data analysis, yielding the plots and results we see in the paper. If other researchers were interested in conducting this same user study, I believe the materials provided here would help them achieve that.

Including the Jupyter notebooks in PDF format with the pre-computed cells made it easy to follow the analysis process in detail. In addition, after deploying the Python environment, it was easy to repeat the same results.

There were a couple of minor details I summarize next:

Ambiguous sentence in the paper. In Section 6.2 the paper states: "'We determined that the data was not normally distributed, as is required for a parametric statistical test"'. This sentence is ambiguous and made me initially think that a parametric test was used for non-gaussian data, which would have been wrong. The analysis, in fact, uses a non-parametric test which is correct when the data distribution is not gaussian. The paper also states the use of a non-parametric test later in the text.

Repeating data analysis results. I summarized some of the minor issues in deploying Python environments above and suggested an alternative.

4.2 Results

I could repeat the plots presented in the paper as well as the exact data analysis.

5 SUMMARY

The materials provided detail the data analysis and help build confidence in the results of the paper. The instructions included are sufficient for other researchers to recreate the user study.

[1] At least to conclude statistical significance using the same significance threshold.