CS263 Final project

Data Science Performance Comparison and Benchmarks for Python, MATLAB, and R

Katherine Turnlund
dept. Computer Science
kturnlund@ucsb.edu

Jaber Daneshamooz
dept. Computer science
jaber@ucsb.edu

***Abstract*—put abstract here**

***Index Terms*—Data science, performance analysis, python, MATLAB, R**

I. Introduction

In this project, we compared the performance of data science algorithms between MATLAB, R, and Python. All of these three programming languages are interpreted languages ans widely used for data analysis; Therefore, our comparison will show which language excels and which one struggles. Since there were no good bench marking available for theses languages in data science field, we implemented some programs to test that. These programs are the ones which are frequently used in data science field.

II. Gap of knowledge

No data science bench marking

III. Overview of programming languages

A. Python

Python is a dynamically typed language and has interpreter rather than compiler. The python byte-code has 113 opcodes(42 with arguments/operands and 71 without). CPython is the reference implementation of the Python programming language. Written in C and Python, CPython is the default and most widely used implementation of the Python language. CPython can be defined as both an interpreter and a compiler as it compiles Python code into bytecode before interpreting it.

Cpython compiles Python code into bytecode before interpreting it makes use of a global interpreter lock (GIL) on each CPython interpreter process which means that it is single threaded.The GIL makes python unsuitable for CPU-intensive algorithms which run across multiple cores. Cpython has a fat bytecode and more work is done in handler for each bytecode. Decode and dispatch

B. R

The R programming language is strongly but dynamically typed. It is functional and interpreted and therefore, not compiled. R is popular amongst data scientists, because there are (free) packages with which statistical calculations (such as matrix calculations or descriptive statistics) can be performed.

C. MATLAB

Matlab is loosely and weakly typed as well as dynamically typed programming language. The aformentioned properties can cause problem when re-using or changing the code. Matlab is a scripted rather than compiled language and uses an interpreter. Matlab code is interpreted and optimized by JIT(Just In Time) accelerator which removes the redundant code and re-orders the commands.

One of the reasons that Matlab is slower than some of the programming languages is that it performs extra operations. For example, it checks such whether the things passed to the functions are appropriate or not. Also, sometimes Matlab performs some analysis to determine the best internal method for accomplishing the operation. This operations consume CPU cycles and make Matlab slower than its counterparts.

Matlab has a compiler as an extra toolbox for sharing. The compiler allows the client to share the code without the need of buying MATLAB for destination. Using Matlab compiler, you can deliver the Matlab program as a standalone application, web-app, docker image, excel add-in etc [1].

D. Java

Provided in the course slides and we do not need to present them here

IV. Code samples

A. Criteria and Variation

We implemented different progra

B. Linear Regression

code in java and python and matlab and R
salary based on years of experience
put the generated data on git hub
plotting vs not plotting
mention java libraries for plotting

C. Matrix Multiplication

code in java and python and matlab and R

D. Data slicing

pandas

E. Basic stats

fibonacci
basic objects

V. Benchmark

A. Strategy

Run each of the toy programs 10 times??? on the same system Make sure the system has similar background workload while running all of the tests with different languages Run all of the tests at the same time !!! Analyze the derived data Plot the corresponding graphs for better analysis Leverage multicore and single core systems for test

Running all of the codes on the same device (same hardware resources) Making sure there is no heavy background process is running on the computer during the execution of benchmarks Running each code 10 times and taking the average, max and min Running the benchmark couple of times in different times of the day to get more accurate data Running all of the codes simultaniously to see which one would be the winner Considering which programming lang can get the resources faster and run faster Compare two by two: eg Java vs Python , Java vs R Compare the benchmark of already compiled codes with the interpreted ones

B. statistics

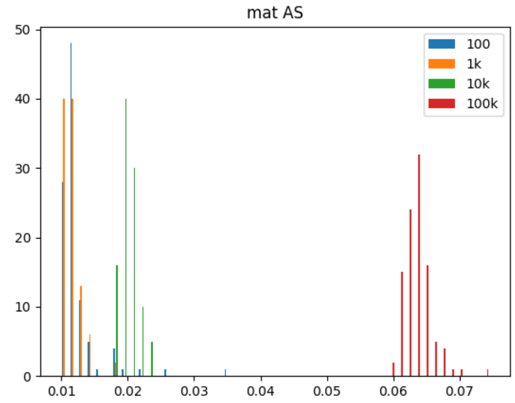

Fig. 1. Matlab Array Slicing

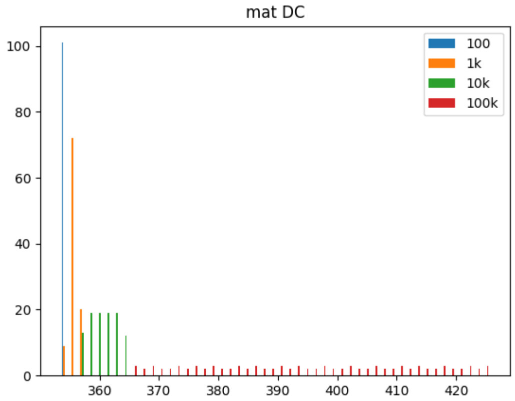

Fig. 2. Matlab Data Cleaning

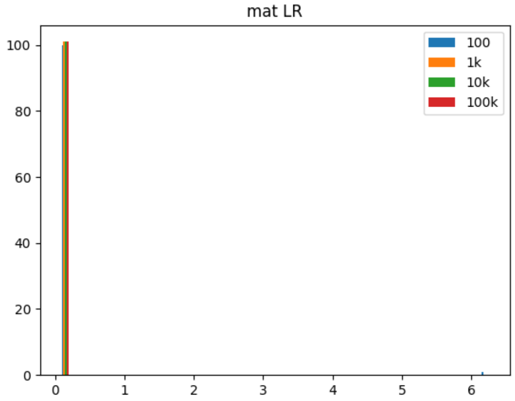

Fig. 3. Matlab Linear Regression

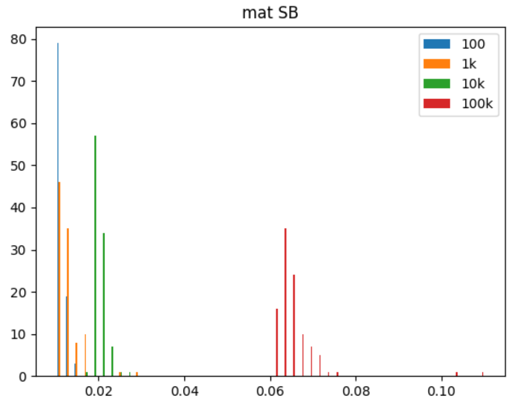

Fig. 4. Matlab Basic Stats

1) Matlab:

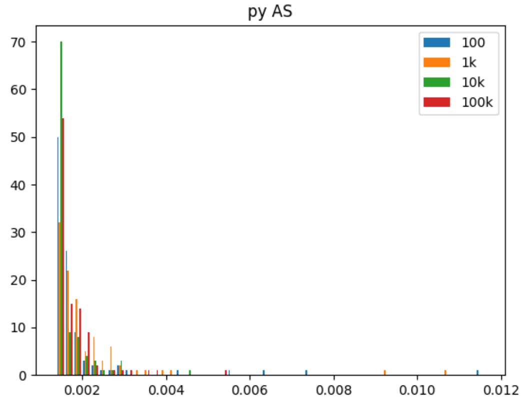

Fig. 5. Python Array Slicing

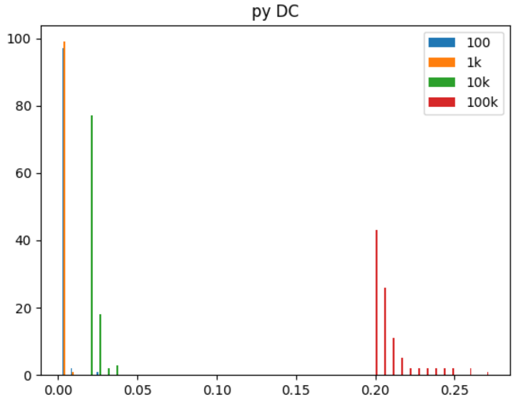

Fig. 6. Python Data Cleaning

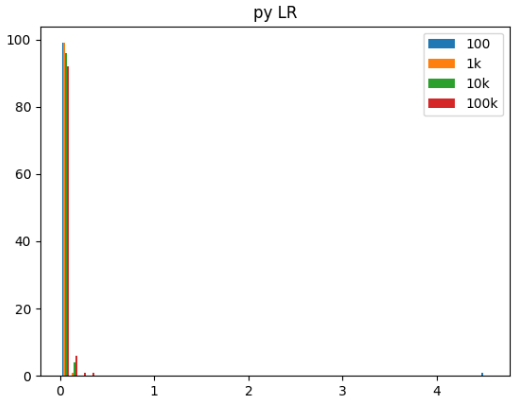

Fig. 7. Python Linear Regression

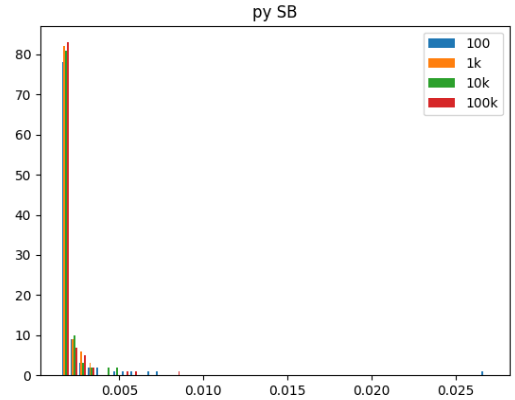

Fig. 8. Python Basic Stats

2) Python:

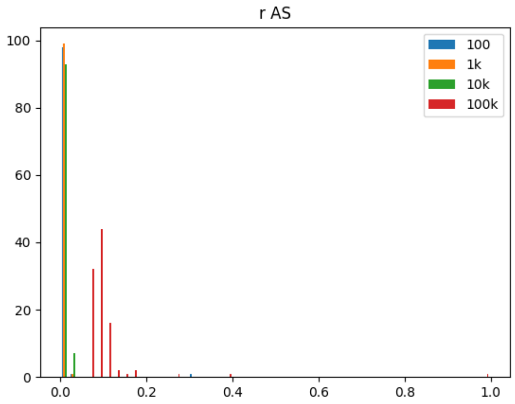

Fig. 9. R Array Slicing

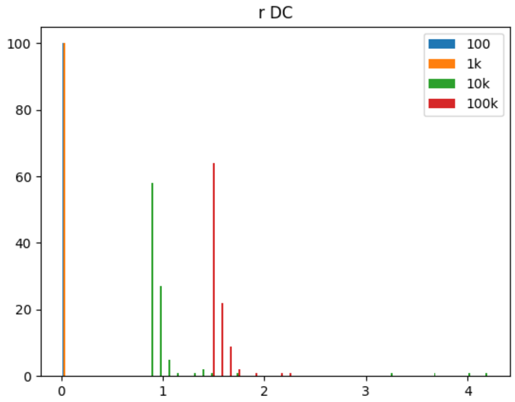

Fig. 10. R Data Cleaning

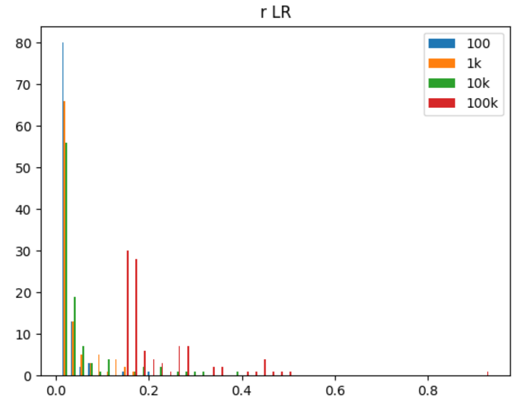

Fig. 11. R Linear Regression

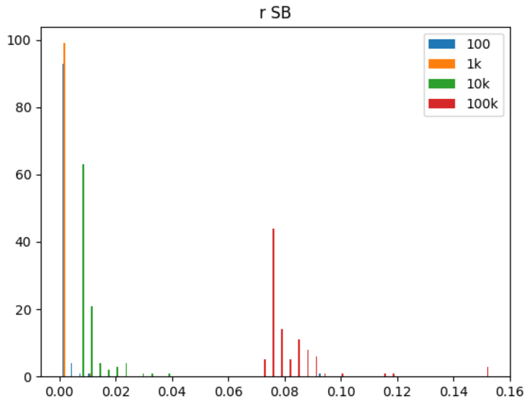

Fig. 12. R Basic Stats

3) R:

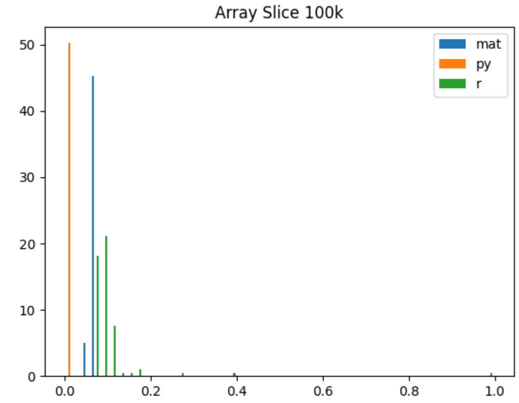

Fig. 13. Array Slicing Comparison

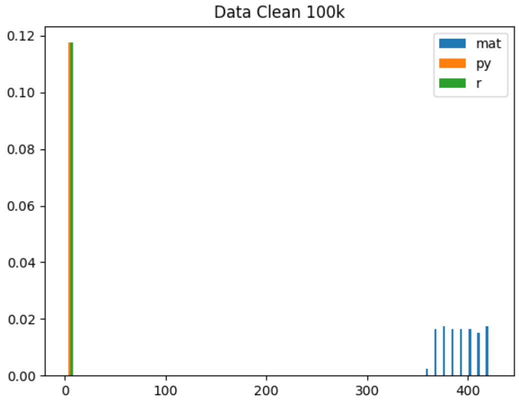

Fig. 14. Data Cleaning Comparison

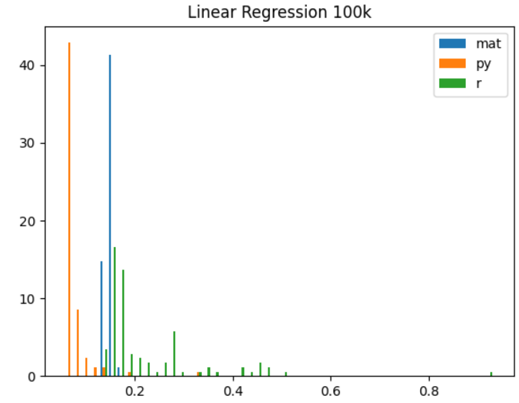

Fig. 15. Linear Regression Comparison

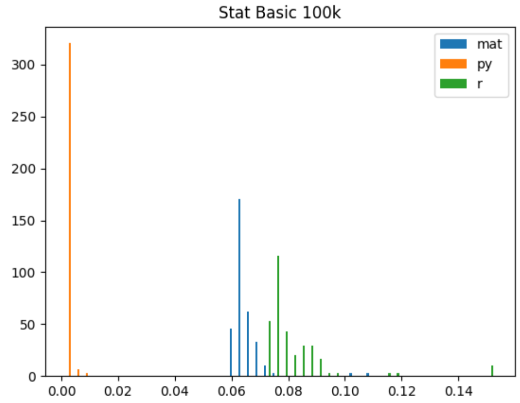

Fig. 16. Basic Stats Comparison

4) Comparison:

VI. Comparison of user experience

VII. Challenges and accomplishments

VIII. Future works

IX. Acknowledgment

Put ack here
// end of large font

References

[1] MATLAB Compiler

[2] Cpython

[3] R Documentation

[4] Benchmarksgame

[5] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.