ECE/CS 498 DS HW 1 Spring 2020

Name: Chuhao Feng

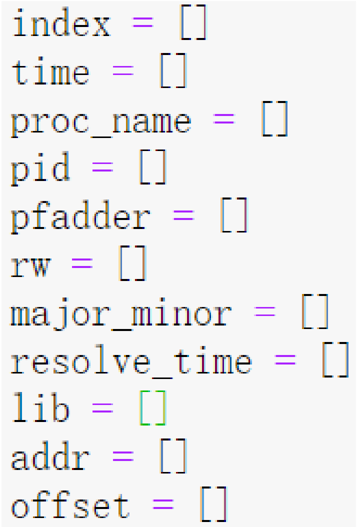
Netid: chuhaof2

Registration Status: registered

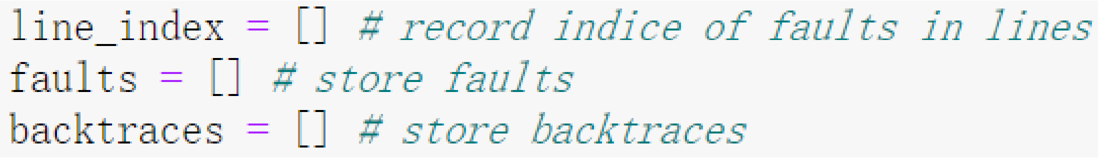
A. Data Structure to Parse Raw Log File

- Provide a (i) diagram and (ii) brief explanation of the data structure you used to parse the raw log file
- Basically, I used lists to parse the raw log file. First, I used line_index, faults, and backtraces to store preliminary data from the raw log file. Second, I identified the data in these three lists and stored the identified data into the corresponding 11 lists listed to the right. Third, I used these 11 lists to create data_dict. Finally, I used data_dict to make the required data frame.

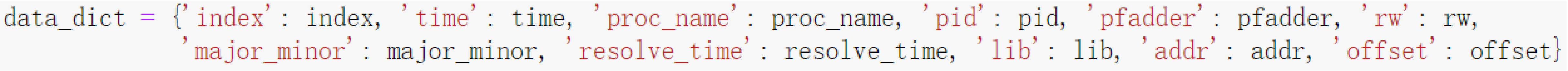
  (refer to .ipynb for detail)



```
index = []
time = []
proc_name = []
pid = []
pfadder = []
rw = []
major_minor = []
resolve_time = []
lib = []
addr = []
offset = []
```



```
line_index = [] # record indice of faults in lines
faults = [] # store faults
backtraces = [] # store backtraces
```



```
data_dict = {'index': index, 'time': time, 'proc_name': proc_name, 'pid': pid, 'pfadder': pfadder, 'rw': rw,
             'major_minor': major_minor, 'resolve_time': resolve_time, 'lib': lib, 'addr': addr, 'offset': offset}
```

B.a. Time Range Covered By Data

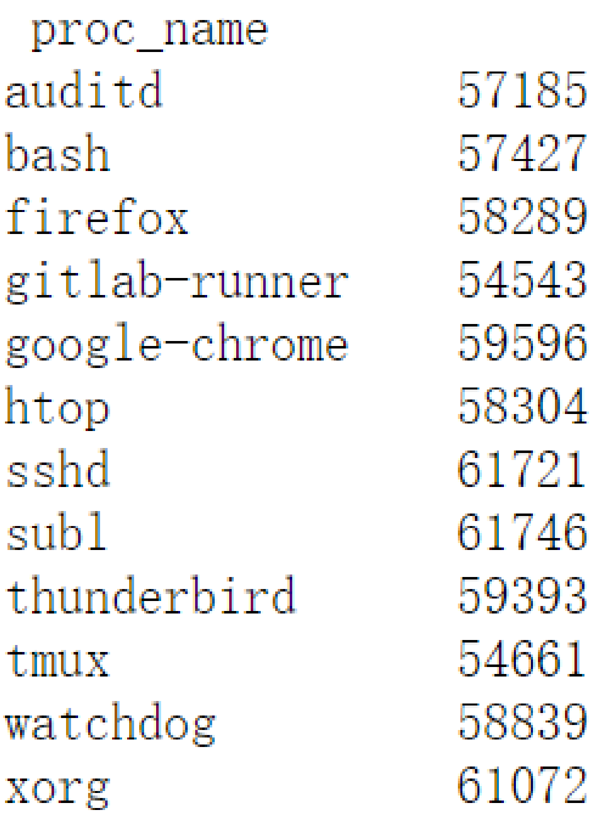
- Start Time: 2017-10-01 00:01:09.251000
- End Time: 2018-01-07 18:59:50.839000
- Total Duration: 98 days 18:58:41.588000

B.b. Unique Processes

- Include
  - The number of unique processes: 12
  - The name of each process:

  auditd, bash, firefox, gitlab-runner, google-chrome, htop, sshd, subl, thunderbird, tmux, watchdog, xorg

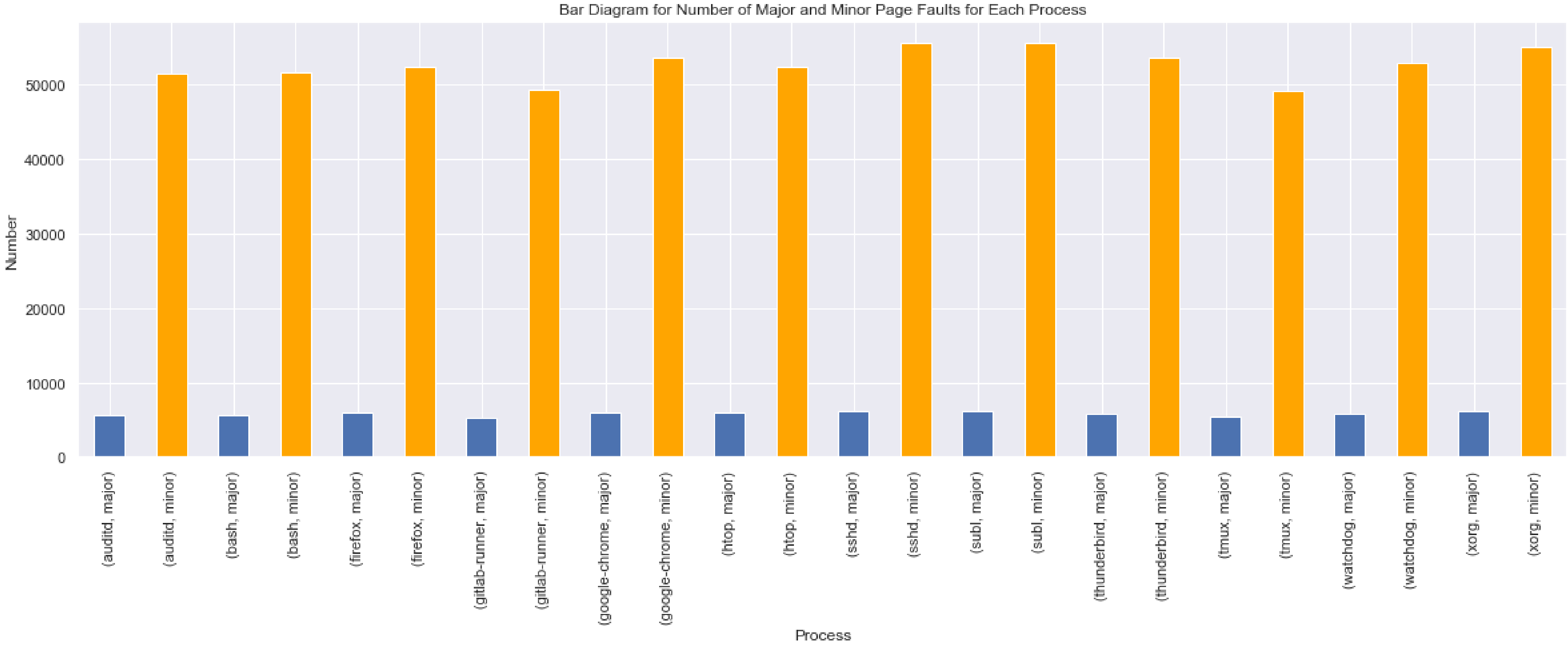
  - The number of times each process was executed



```
 proc_name
auditd           57185
bash             57427
firefox          58289
gitlab-runner    54543
google-chrome    59596
htop             58304
sshd             61721
subl             61746
thunderbird      59393
tmux             54661
watchdog         58839
xorg             61072
```

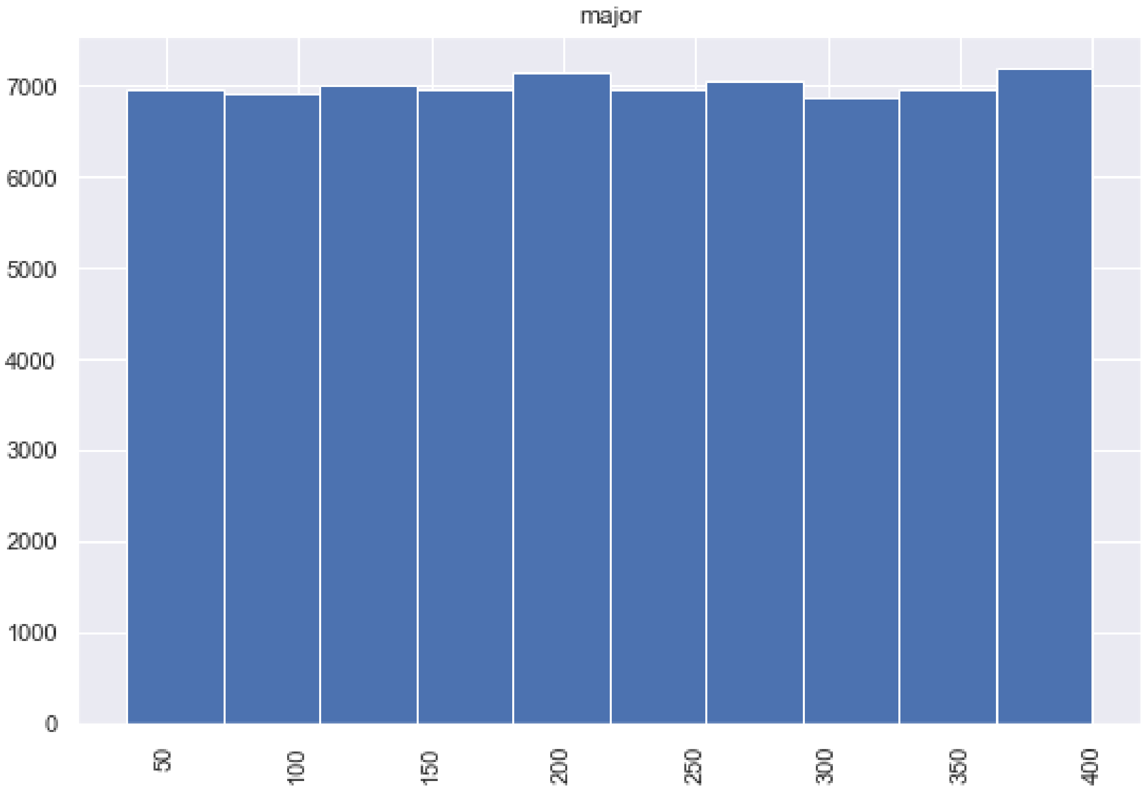
B.c. Major and Minor Page Faults

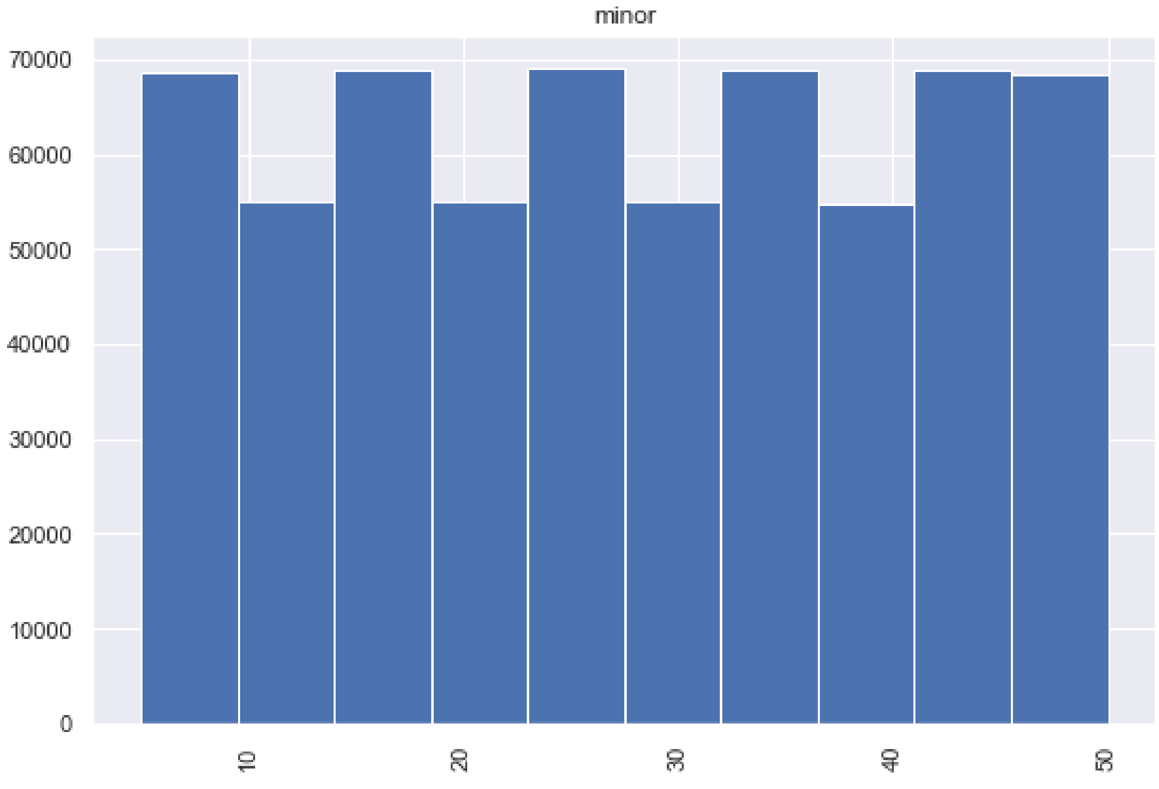
- Include a bar chart showing the number of major and minor page faults for each process (refer to .ipynb for a more legible chart)
- Remember to include axes labels and a title!



B.d. Time to Resolve Page Faults

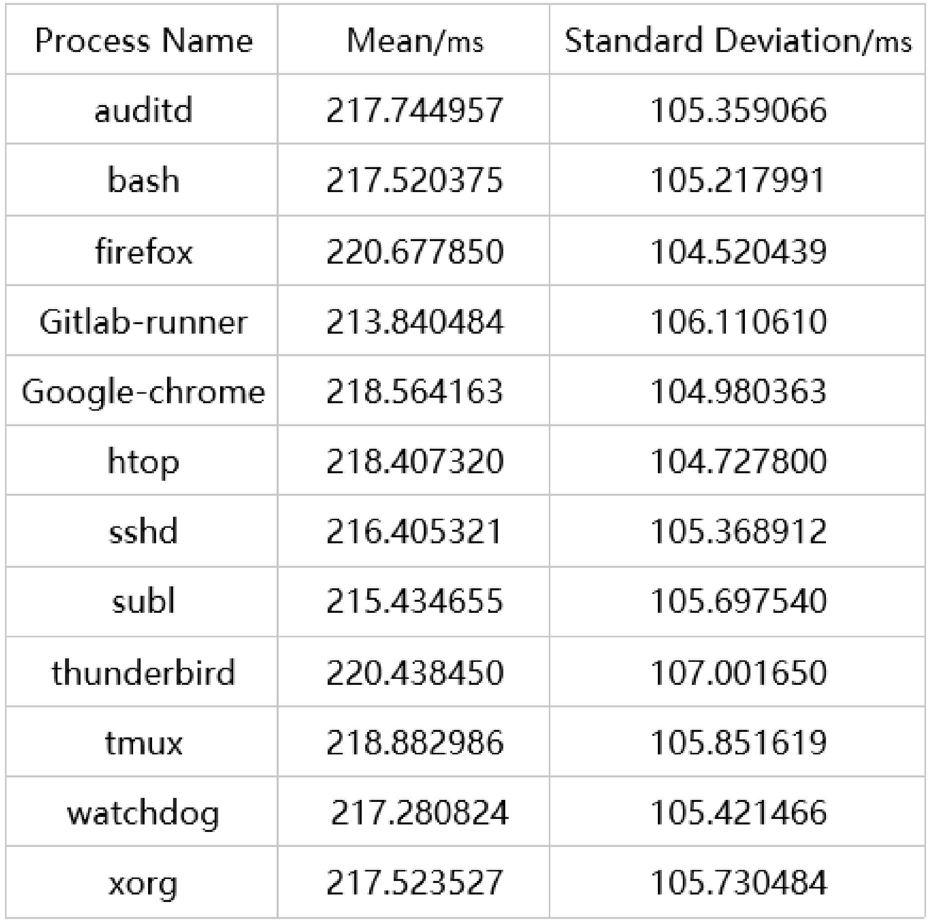
- Include
  - Histogram of time to resolve minor page faults
  - Histogram of time to resolve major page faults





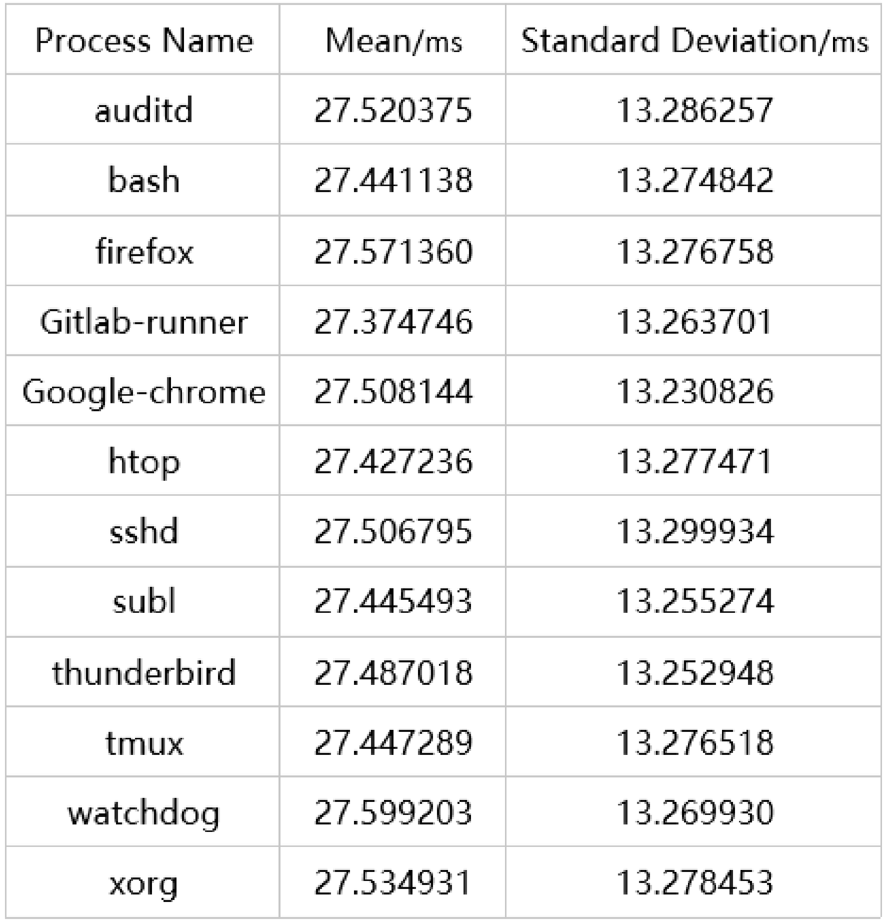
- Table with mean and standard deviations of times to resolve page faults for each process, separated by fault severity (i.e. major or minor)

Major



Process Name	Mean/ms	Standard Deviation/ms
auditd	217.744957	105.359066
bash	217.520375	105.217991
firefox	220.677850	104.520439
Gitlab-runner	213.840484	106.110610
Google-chrome	218.564163	104.980363
htop	218.407320	104.727800
sshd	216.405321	105.368912
subl	215.434655	105.697540
thunderbird	220.438450	107.001650
tmux	218.882986	105.851619
watchdog	217.280824	105.421466
xorg	217.523527	105.730484

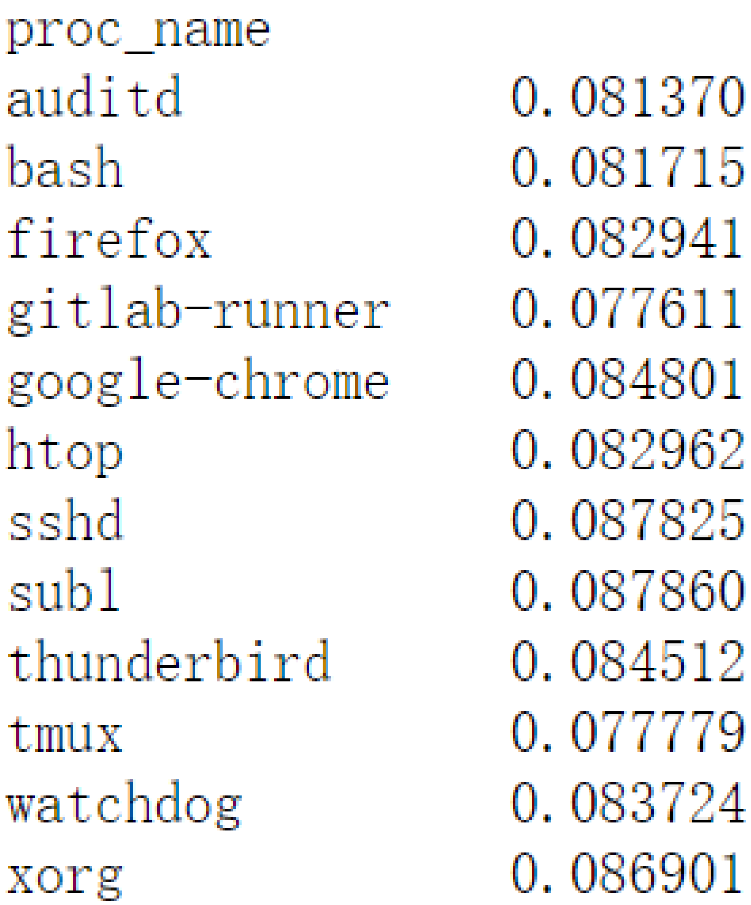
Minor



Process Name	Mean/ms	Standard Deviation/ms
auditd	27.520375	13.286257
bash	27.441138	13.274842
firefox	27.571360	13.276758
Gitlab-runner	27.374746	13.263701
Google-chrome	27.508144	13.230826
htop	27.427236	13.277471
sshd	27.506795	13.299934
subl	27.445493	13.255274
thunderbird	27.487018	13.252948
tmux	27.447289	13.276518
watchdog	27.599203	13.269930
xorg	27.534931	13.278453

C.a. Class Priors

- List the priors for all the classes



```
proc_name
auditd           0.081370
bash             0.081715
firefox          0.082941
gitlab-runner    0.077611
google-chrome    0.084801
htop             0.082962
sshd             0.087825
subl             0.087860
thunderbird      0.084512
tmux             0.077779
watchdog         0.083724
xorg             0.086901
```

C.b. – C.c. : Predictions

- Given that the page fault was major, which process was it most likely caused by? (refer to .ipynb for detail explanation)
- The page fault is most likely caused by process **subl**.

- Given that the page fault was from a read access, which process was it most likely caused by? (Refer to .ipynb for detail explanation)
- The page fault is most likely caused by process **subl**.

C.d. Appropriate Model

- In 2 sentences or less, explain which model taught in class could be used for classifying the process given information about the fault’s (i) severity and (ii) access type.
- Naïve Bayes model could be used for classifying the process given information about two distinct features, because it is reasonable and intuitive to assume these two features are independent of each other given certain process type. Besides, Naïve Bayes model can be used as a classifier.