Next-Generation sequencing creates many short sequence reads that align to multiple locations in the genome. An important metric often calculated is genome coverage – how well was our genome covered by sequencing reads, i.e. the “read depth”. The objective of this assessment is to determine some frequently calculated coverage metrics.

There are three files containing read depths for three patients in a sequencing experiment, where each row of data is formatted like so:

|  |  |  |
| --- | --- | --- |
| Chromosome | Position | Observed Read Depth |

Included in all three files are depth values for positions 1 to 1,000,000 for chr1, chr2, and chr3.

Not all positions are contained in each file – a missing position means that the sample has no coverage at that position (0 reads aligned to that position). As such, positions that have a positive depth value in one file may be missing in other files.

Write a solution using Python which answers the following questions as succinctly, efficiently, and clearly as possible, and which leaves the input files unmodified. You may utilize any commonly used open source packages you like. Please note that if you include any kind of ‘best practices’ code (e.g. docstrings, comments) these lines won’t count against you and are in fact welcomed:

1. How many positions are unique to the file sample2\_depths.txt?
2. How many positions are in common between all three depth files?
3. What is the average read depth for each chromosome in the experiment (not the chromosome averages for each individual sample)? Consider each chromosome to have a total of 1M positions.
4. Which position has the largest average depth?

Please complete your Python solution within 24 hours. Good luck!