For this program, the basic idea is below:

For every thread, we read the certain part of the text file. The position is decided in this way: first we calculate how large this text file is, then we use this number divided by the number of threads we use. Then it’s easy to calculate which thread read which part of this file.

After reading the file, we store the tick into vector. Then we sort this vector use sort function with the comparing method we construct. In this way, every vector is well sorted by time.

Then we try to decide which entry is noise. There are basically three aspect: one, if the price is below 0, then it’s noise; two, if two timestamp is the same, then we decide the second one as noise; three, if the price is 3 deviation away from the mean, we recognize it as the noise.

At the end, we output the tick into two files.