数据分析表面上看起来简单,其实需要花大量的时间。首先，有大量数据缺失，每个 CSV 文件中，具体推文数量也不一样。如果需要从一个整体的角度进行清理分析，就需要把几套数据按一个 Key 进行 merge,这里使用的是 tweet id 来把两张数据表进行合并。同时利用肉眼观察，能发现 time stamp 里的时间显得杂乱，不利落直观，另一方如果利用 info()可以发现 time stamp 的数据类型也不正确，一般会使用 datetime 作为时间的数据类型。狗狗地位数据也有大量缺失，而且也没有必要将四种地位分到四个列，造成存储空间浪费。对此，我们可以将四列合并为一列（列名可以是具有概括性的词语）.

我也意识到，此套数据还有相当多其他方面的缺陷，也是一个非常好的熟悉各种 panda library 的过程。在做数据分析时，需要明确目的，才能避免浪费时间。由于对各种库的不熟悉，不了解，也会造成分析过程效率十分低下