Wrangle Report

本项目的数据整理过程一共分为以下三个大部分，分别是数据的收集，数据质量和整洁度的评估，数据的清理。第一部分数据的收集共有三个 csv 文件，其中有两个是项目自行提供的，分别是来自'We Rate Dogs' 的五千条 twitter 的基本信息和相应的图像预测结果，最后一个数据集是每一条 twitter 附加的信息，需要通过 twitter 的 API 去自行获取。
第二部分是对数据质量和整洁度的评估，首先是目测评估，然后是进一步编程评估。目测评估中会发现一些简单直观的问题，如空值的存在以及表冗余，其分别属于质量问题和整洁度问题。编程评估会发现一些更具体的问题，如时间类型的转换，时间的过滤，删除掉含有空值的无效的列以及删除掉那些转发或者回复的 twitter（因为我们只关注原始的 twitter）等等，这些都是质量问题，也有诸如列冗余的问题，一个变量只应该占一列。
第三部分就是利用 Pandas 库对以上所述的问题进行一一清理解决。
在完成以上三步工作之后，需要对最终版本的数据进行简单的分析，这里是做出了三个总结结论和一个可视化图形。