wrangle 项目数据整理报告

整理 (以及分析和可视化) 的数据集是推特用户 @dog_rates 的档案，推特昵称为 WeRateDogs。WeRateDogs 是一个推特主，他以诙谐幽默的方式对人们的宠物狗评分。这些评分通常以 10 作为分母。但是分子则一般大于 10：11/10、12/10、13/10 等等。为什么会有这样的评分？因为 "They're good dogs Brent." WeRateDogs 拥有四百多万关注者，曾受到国际媒体的报道。

WeRateDogs 下载了他们的推特档案，这个档案是基本的推特数据（推特 ID、时间戳、推特文本等），包含了截止到 2017 年 4 月 1 日的 5000 多条推特。这份推特档案很棒，但是只包含基本的推特信息。要达到 "Wow!" 的效果，在分析和可视化前，还需要收集额外的数据、然后进行评估和清洗。

档案中有一列包含每个推特的文本，提供的数据文件用这一列数据提取了评分、狗的名字和“地位”（即 doggo、floofer、pupper 和 puppo）——这使数据得以“完善”。在这 5000 多条中，只筛选出了 2356 条包含评分的推特数据。从数据评估结果看，提取的评分和狗的名字比较完整，狗的地位只有少量数据，四个地位数据总共只有 400 条，与总的推特数据 2356 差距太大，不具有统计意义，做删除处理。

课程提供了对推特中图片进行识别后的处理结果，对出现在每个推特中狗的品种（或其他物体、动物等）进行了预测，这些数据需要通过提供的 URL 来进行编程下载。我对下载后的数据只选取了可以被识别为狗最可信的识别结果，以及识别为其他动物的最可信的结果。然后把这些数据合并到了推特档案数据中。

课程还提供了推特的补充数据，里面包含大量内容，由于对字段

的含义不了解，只选取了 tweet_id 和 retweet_count，favorite_count 这三类数据，最终合并到主数据集推特档案中。

评估数据时发现 rating_numerator 和 rating_denominator 这两类数据的最大值与平均值偏离太大，起初认定有可能是异常值。在课程的关键要点中有说明“如果分子评级超过分母评级，不需要进行清洗。这个 特殊评分系统 是 WeRateDogs 人气度较高的主要原因。（同样，也不需要删除分子小于分母的数据）”，根据此说明推测异常的最大值是有可能的。另外，视觉评估时发现，excel 中显示 rating_numerator 和 rating_denominator 混入了前边 text 列的内容，通过编程证明是 excel 读取 csv 文件问题,数据正确。

在合并三张原始数据表时，由于图像识别和推特补充数据中缺少部分推特档案的对应信息，合并后的数据自动转变成了 object 类型，缺失的数据为 NaN，不方便后续的统计分析，所以合并后设置了默认值，将类型修改回了原始类型。

经过评估和清理后，我得到了一份比较整洁和清晰的数据。使用这份数据分析了发布信息最多的平台，识别率高的狗的品种，和转发最大最受喜欢的狗品种。